Highlights of (bio-)chemical tools and visualization software for computational science

Dubbeldam, David; Vreede, Jocelyne; Vlugt, Thijs JH; Calero, Sofia

DOI
10.1016/j.coche.2019.02.001

Publication date
2019

Document Version
Accepted author manuscript

Published in
Current Opinion in Chemical Engineering

Highlights of (Bio-)Chemical Tools and Visualization Software for Computational Science

David Dubbeldam

Van't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands

Jocelyne Vreede

Van't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands

Thijs J. H. Vlugt

Engineering Thermodynamics, Process & Energy Department, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Leeghwaterstraat 39, 2628CB Delft, The Netherlands

Sofia Calero

Department of Physical, Chemical and Natural Systems, Sevilla 41013, University Pablo de Olavide, Spain

Abstract

Computational chemistry uses computer simulation to assist in solving chemical problems. Typical workflows of computational chemists include the use of dozens of utilities. 3D modeling programs are powerful tools that help researchers visualize their work and create illustrative graphics. In this review, we describe and highlight tools and visualization packages that are commonly used in the field of (bio-)chemistry and material science.

Keywords: Material Science, Bio-chemistry, Computational utilities, software

Email address: d.dubbeldam@uva.nl (David Dubbeldam)

1. Introduction

Molecular simulation is a powerful tool to conduct "in-silico" experiments. At the atomic level there are quantum-mechanical packages that compute atomic properties of a few atoms very accurately using coupled-cluster approaches. Density Functional Theory (DFT) is currently applicable to hundreds of atoms, while a classical formulation can handle trillions of atoms. Using meso-scopic and hybrid modeling larger systems and longer time-scales can be reached. Continuum mechanics like computational fluid dynamics (CFD) handles the largest space and longest time-scales, and is based on partial differential equations. Many systems require a multiscale approach, as macroscale models do not provide atomic insight, while the microscale models are computationally demanding.

There are hundreds of chemical software packages available, at many different scales of resolution. Examples of popular and efficient parallel Molecular Dynamics (MD) codes are LAMMPS [1], OPENMM [2], and GROMACS [3]. We refer to table 1 for lists of software packages on the various computational topics, as available on wikipedia. Pirhadi et al. provided a topic perspective on open source molecular modeling [4]*. Kozlikova et al. reviewed the state of the art of visualization of biomolecular structures [5]. In this review, we will not focus on these software packages, but will highlight utilities that are used to setup input for software packages, to convert file formats and force fields, and to analyse and visualize the results. We will mainly discuss packages that are used within our own groups with the aim to make newcomers to the field of computational chemistry aware of the existence and value of these packages.

	https://en.wikipedia.org/wiki/
academic databases	List_of_academic_databases_and_search_engines
algebra systems	List_of_computer_algebra_systems
analysis software	List_of_numerical_analysis_software
bioinformatics	List_of_open-source_bioinformatics_software
chemical processes	List_of_chemical_process_simulators
cheminformatics	Cheminformatics_toolkits
computer simulation	List_of_computer_simulation_software
deep learning	Comparison_of_deep_learning_software
finite element	List_of_finite_element_software_packages
quantum chemistry	List_of_quantum_chemistry_and_solid-state_physics_software
modeling on GPU	Molecular_modeling_on_GPUs
molecule editor	Molecule_editor
molecular design	Molecular_design_software
molecular mechanics	Comparison_of_software_for_molecular_mechanics_modeling
Monte Carlo	List_of_software_for_Monte_Carlo_molecular_modeling
nanostructures	List_of_software_for_nanostructures_modeling
nucleic acid	Comparison_of_nucleic_acid_simulation_software
numerical analysis	Comparison_of_numerical_analysis_software
optimization software	List_of_optimization_software
plotting software	List_of_information_graphics_software
protein-ligand docking	List_of_protein-ligand_docking_software
protein structure prediction	List_of_protein_structure_prediction_software
SMILES related	Simplified_molecular-input_line-entry_system
statistics	Comparison_of_statistical_packages
visualization	List_of_molecular_graphics_systems

Table 1: Wikipedia entries on list of software packages.

2. Materials, structures and molecules

2.1. Types of molecules and materials

There are many types of materials, e.g. biomaterials, ceramics, composites, metals, nanoporous materials, (porous) polymers, semiconductors, and smart materials. In such systems, the atoms and/or molecules are closely packed and have a natural resistance to change of shape/volume. In crystalline materials the atoms are arranged in a regular repeating three-dimensional array, while more or less randomly arranged solids are called amorphous.

Macromolecules are very large polymeric molecules such as proteins, carbohydrates, nucleic acids, and polyphenols, or large non-polymeric molecules such as lipids and macrocycles. Most macromolecules are polymers, which are long chains of subunits called monomers. Molecules much smaller than macromolecules or molecules of low molecular weight are called micromolecules. Proteins are an extremely important group of macromolecules made up of just 20 different amino acids. Amino acids have a chiral carbon attached to a hydrogen, an amino group, a carboxyl group and a rest group, that varies with amino acid type. The amino acids in proteins are connected via peptide bonds, which form the main chain, with the rest groups as side chains. The sequence of amino acids in a protein is called the primary structure of a protein. Hydrogen bonds between peptide bonds within a chain form secondary structure elements, known as the α-helix (repeating coil) and β-sheet (sheets of extended strands) respectively. Interactions between side chains arrange the secondary structure elements into a specific shape known as the tertiary structure. When a protein contains multiple main chains, the arrangement of these chains is called quaternary structure. Similarly, DNA and RNA are composed of (deoxy) ribose nucleotides which contain a phosphate, a pentose sugar, and a nitrogenous base. For both DNA and RNA, four different bases exist in a specific sequence. DNA occurs often in the well-known double-helix structure, while RNA can have many different forms.

With such a variety of materials, structures and molecules, it is no wonder

that there exists a vast array of different file formats to describe and communicate the molecular information. At the very least, the atomic positions and atom type must be present. Often some sort of connectivity information is present to define bonds. Symmetry operations and spacegroup information is needed for crystals. For macromolecules there is also additional information on e.g. sequence number and primary, secondary, tertiary and/or quaternary structure.

2.2. Structure file formats

A common file format for micromolecules is the XYZ-format. A typical XYZ format specifies the molecule geometry by giving the number of atoms with Cartesian coordinates that will be read on the first line, a comment on the second, and the lines of atomic coordinates in the following lines. The units are generally in Ångstroms.

Macromolecules are often reported in the Protein Data Bank (PDB) format. Crystal information can be provided via the unit-cell and space-group records using the Hermann-Mauguin space group symbol. A nice feature of PDBs is that multiple structures can defined using a model serial number. This feature is often exploited in molecular visualizers to create 'movies' (i.e. molecular trajectories). PDB is an 80 column wide line format and hence has limited precision for the atomic positions and charge, as well as a maximum to the number of residues and atom serial numbers.

Crystalline materials, like zeolites and MOFs, are usually reported with a unit cell and a space group in the crystallographic information file (CIF) file format [6]. CIF is a free-format, easily editable archive file constructed to be read by both computer programs and humans. The format is extensible allowing simulation codes to store arbitrary additional property data. Closely related is mmCIF, macromolecular CIF, which is intended as an alternative to the Protein Data Bank (PDB) format. The mmCIF file format is an extension of the CIF representation aimed at solving many of the restrictions of the PDB format.

Most software used in computational chemistry has its own particular file

format, as well as the capability to import and export to other formats. Over a hundred different molecular information file formats are used in chemistry.

2.3. Symmetry and spacegroup information

SgInfo is a comprehensive collection of routines for the handling of space group symmetry. Input Hall symbols are translated into Seitz matrices, which are used to generate the full set of symmetry operations. An online tool to play around with SgInfo to see the Seitz-matrices and spacegroup operators can be found at `http://cci.lbl.gov/sginfo/sginfo-query.cgi`. SgInfo has been superseded by the space group toolbox (sgtbx), which is a part of the open source package Computational Crystallography Toolbox (cctbx) [7]. The source can be found at `https://github.com/cctbx/cctbx_project`. Another program for retrieval of space-group information in several settings and generator-containing space-group symbols is SPGGEN [8].

Spglib is a C-library written for finding crystal symmetry [9]. Avogadro uses spglib to perceive space groups. The library can find symmetry operations, identify the space group type, do Wyckoff position assignment, and find the primitive cell. The source code can be found at `https://github.com/atztogo/spglib`.

Bilbao Crystallographic Server is an open access website offering online crystallographic database and programs aimed at analyzing, calculating and visualizing problems of structural and mathematical crystallography, solid state physics and structural chemistry [10]. The server is accessible at web-address `http://www.cryst.ehu.es`. Automatic Flow for Materials Discovery (AFLOW) is a multi-university research consortium aimed to develop, serve and maintain a plethora of online computational frameworks. AFLOW-SYM is a platform for the complete, automatic and self-consistent symmetry analysis of crystals [11]. The server can be found at `http://aflowlib.org/aflow_online.html`.

2.4. Cheminformatics: SMILES and InChI

Modern chemical notation systems are based on encoding of chemical structures. The simplified molecular-input line-entry system (SMILES) uses human-readable ASCII strings for describing the structure (atoms, bonds, aromaticity,

branching) of chemical species[12]. SMILES are generally obtained by converting a chemical graph to a spanning tree (cycles are broken) and printing the symbol in a depth-first tree traversal. However, since different atoms can be selected as the root, this traversal is non-unique, which was largely overcome by the development of *canonical* SMILES. SMILES arbitrary target specification (SMARTS) is a language that allows you to specify substructures using rules that are straightforward extensions of SMILES. SMIRKS (a hybrid language of SMILES and SMARTS) and SYBYL Line Notation (SLN) allow specification of chemical reactions and wider variety of information. International Chemical Identifier (InChi) is the latest standardized encoding with good canonical serialization of structure, a valence model, stereo centers, and can handle isomers. It is not human-readable however, is more difficult to use for substructure searching and lacks chemical reactions.

One use of SMILES and InChI is to quickly and automatically build molecules. SMILES strings can be imported by most molecule editors for conversion back into two-dimensional drawings or three-dimensional models of the molecules. To interconvert between IUPAC systematic names, InChI strings, SMILES strings, and chemical structures one can use ChemDraw [13]. ChemDraw can interpret SMILES and InChI strings from text into chemical 2D and 3D structures. Cactus is a resolver module available at CIR (Chemical Identifier Resolver): "name patterns". It allows for Google-like searches on a name index of more than 70 million names. This service works as a resolver for different chemical structure identifiers and allows the conversion of a given structure identifier into another representation or structure identifier. It can be used via a web form or a simple URL API. For example, to obtain the SMILES string for DABCO (1,4-diazabicyclo[2.2.2]octane)

```
curl https://cactus.nci.nih.gov/chemical/structure/dabco/smiles -o dabco.smi
```

3D conformations can also be generated with RDKit, a free opensource cheminformatics package.

SMILES and InChi notations of Lewis structures are also particularly useful for constructing databases of molecules that are employed in screening studies. A single-line notation facilitates the storage of molecular structures in a string field, which is supported by any database implementation. Through canonical-

ization, duplicates are easily detected, which facilitates curation of data sets.

3. Utilities

Database searches. Databases of compounds and metadata can be used for screening, either 2D (substructure, similarity) or 3D (shape, pharmacophores). ChemSpider is a free chemical structure database providing fast text and structure search access to over 67 million structures from hundreds of data sources. An extensive list of chemical databases is available at the online chemistry guide `http://www.chemistryguide.org/chemical-databases.html`.

PubChem (`https://pubchem.ncbi.nlm.nih.gov`) is a public repository for information on chemical substances [14], and can be accessed for free through a web user interface. PubChem contains substance descriptions and small molecules with fewer than 1000 atoms and 1000 bonds. It contains its own online molecule editor with SMILES/SMARTS and InChI support that allows the import and export of all common chemical file formats to search for structures and fragments.

Reaxys provides access to Beilstein CrossFire, an online chemical encyclopedia, containing all the important information about more than 7 million organic chemical compounds, from 1771 to the present, including reactions and chemical and physical properties (with all corresponding literature references). Reaxys also provides access to the Gmelin database for inorganic chemistry (very large repository of organometallic and inorganic information), and the Patent Chemistry database.

The Cambridge Structural Database (CSD) is a comprehensive and up-to-date database of crystal structural with over 950,000 curated entries. The hypothetical zeolite database [15] and the atlas of prospective zeolite structures (`http://www.hypotheticalzeolites.net`) contain million of structures. Other databases are the CoRE MOF database [16, 17] and zeolites IZA structures [18]. MOF Lab is an educational and research tool that provides an online platform to visualize Metal-Organic Frameworks and calculate their physical

properties, available at `http://mausdin.github.io/MOFsite/mofPage.html`.

For proteins and DNA/RNA many databases exists that each provide different aspects of their structure. A sequence of amino acids or nucleotides can be matched using BLAST[19] (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`) against a database of sequences of known proteins or genes, to for instance determine if an unknown mouse gene also occurs in the human genome. The Protein Data Bank (`http://www.rcsb.org`)[20] contains all known 3D structures of proteins and nucleic acids, as determined by X-ray crystallography, NMR spectroscopy, and for an increasing number electron microscopy [20]. When starting with a target amino acid sequence (or gene), a common procedure is to first find out to which protein the sequence belongs using BLAST and then search the PDB for structural information on that protein. As there are far less 3D structures of biological marcomolecules available, the structures in the PDB are used to obtain structural information at different levels on a target sequence. If the target sequence is very similar to the sequence of a structure in the PDB, that structure can be used as a template for the structure of the target sequence, a procedure called homology modeling or comparative modeling. Commonly used programs for homology modeling are Modeller [21] and Swissmodel [22]. By using PSI-BLAST [23], amino acid sequences are matched based on potential structural similarity, providing 3D structural information for sequences which is not necessarily very similar at sequence level.

Molecular drawing. Two highly popular commercial packages are ChemDoodle and ChemDraw. ChemDoodle began as a quality and affordable chemical sketcher, but was later extended to a scientific visualization platform. ChemDoodle has an interface to search directly the ChemExper Chemical Directory from within the program. ChemDraw is a simple-to-use program that allows to draw intuitively and efficiently simple two-dimensional representations of organic molecules. ChemDraw [13], along with Chem3D and ChemFinder, is part of the ChemOffice suite of programs and is available for macOS and windows.

JSME a free molecular editor in javascript [24] Popular windows freeware

sketchers are ChemSketch and MarvinSketch. A very nice online 2D sketcher is `http://molview.org`.

Format conversion. Open Babel is a great utility to convert the format of structures, with over 110 chemical file formats supported [25]**. For example, to obtain a 3D structure for DABCO (1,4-diazabicyclo[2.2.2]octane) one can first obtain the SMILES, and then use Open Babel to convert the SMILES to a three-dimensional molecule in XYZ-format:

```
curl https://cactus.nci.nih.gov/chemical/structure/dabco/smiles -o dabco.smi
babel -ismi dabco.smi -oxyz dabco.xyz --gen3D
```

As an example for an online conversion-tool, see:

`https://www.webqc.org/molecularformatsconverter.php`

Plotting. Examples of nice plotting utilities on linux (and macOS) are gnuplot and xmgrace. Gnuplot can be run interactively, or from script files. Script files are simple ascii files that have commands written out just as you would enter them interactively. Popular programs on windows are Graphpad Prism (also for macOS), Origin Pro, SPSS, and Sigmaplot. Most of these interact very well with data from excel spreadsheets. Computer algebra systems like matlab, maple, and mathematica also provide rich plotting functionality.

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. The Notebook has support for over 40 programming languages, including Python. Matplotlib is an excellent 2D and 3D graphics library for generating scientific figures from within python.

Force field, atom typing, and conversion. Quantum mechanical packages require a working directory with several short ascii-based input-files present. The details differ and depend on the actual program. VASP for example, requires four files called INCAR, POSCAR, KPOINTS and POTCAR. The POTCAR has to be created with pseudopotentials for each atomic species, KPOINTS specifies

the k-points. The POSCAR file contains the lattice geometry and the ionic positions and the ordering must be consistent with the POTCAR file. The INCAR file is the central input file of VASP to specify the simulation type, energy cutoff, etc. Programs like Vesta and iRASPA are able to create POSCAR files. VASPKIT (`http://vaspkit.sourceforge.net`) is a post-processing package for VASP. C2x is a tool for visualisation and input preparation for Castep and other electronic structure codes [26].

Compare to QM code, classical codes are significantly more challenging to setup. Fortunately, many programs like GROMACS, Tinker, Materials Studio, etc, allow you specify generic force field such as CHARMM, AMBER, UFF, etc. This process involves the "typing" of atoms from their chemical element into the force field name. Elements in a different chemical environment have different types, usually based on their neighbors or aromaticity.

The CHARMM General Force Field (CGenFF) program is a product of the discontinued ParamChem project. The program performs atom typing and assignment of parameters and charges by analogy in a fully automated fashion. For AMBER, parameters for molecules can be obtained from the General AMBER Force Field (GAFF)[27, 28] using the tool antechamber (free in AmberTools). The force field conversion from one MD program to another one is exhausting and error-prone. A generic tool for the conversion in both direction for favorite MD programs AMBER, CHARMM, DL POLY, GROMACS, and LAMMPS is ForConX [29].

On online automated topology builder is `https://atb.uq.edu.au`. This site provides access to classical force fields in formats compatible with GROMACS, GROMOS and LAMMPS simulation packages and a GROMOS to AMBER topology file converter. Arpeggio is an online web server for calculating and visualising interatomic interactions in protein structures [30].

Trajectory analysis. The python library MDTraj [31] allows users to manipulate MD trajectories via the implementation of extensive analysis routines. With MDTraj almost any MD format can be read in and written out, perform very fast

analysis, such as RMSD or distance calculations, secondary structure assignment in proteins and computations of experimental observables.

HTMoL is a full-stack solution for remote access, visualization, and analysis of molecular dynamics trajectory data [32]. On online web utility for viewing and sharing molecular dynamics simulations is MDsrv [33].

Equations of state and thermodynamics properties. The Reference Fluid Thermodynamic and Transport Properties (REFPROP) database by NIST, available commercially at `https://www.nist.gov/srd/refprop`, consists of a collection of models and equations of state to describe thermodynamic properties of pure component and mixtures [34]. The following properties are available: temperature, pressure, density, energy, enthalpy, entropy, heat capacity at constant volume and pressure, speed of sound, compressibility factor, Joule Thomson coefficient, 2nd and 3rd virial coefficients, Helmholtz energy, Gibbs energy, heat of vaporization, fugacity, fugacity coefficient, chemical potential, thermal conductivity, viscosity, kinematic viscosity, thermal diffusivity, Prandtl number, surface tension, dielectric constant, gross and net heating values, isothermal compressibility, volume expansivity, isentropic coefficient, adiabatic compressibility, specific heat input, exergy, Gruneisen, critical flow factor, excess values, and others. It is widely used in industry and academics. The latest version runs on Linux, macOS, and Windows.

HSC Chemistry is an advanced software package for thermodynamic and mineral processing calculations. It contains modules for thermodynamic data (thermochemical database), phase equilibria, thermodynamic properties, process simulations using flowsheets, dynamic process simulations, and reaction equilibrium compositions. It also contains a module to perform exergy calculations, to find the lost work in a process. This is a measure for the efficiency of usable energy in the process.

Machine learning. Atomistic Machine-learning Package (Amp) is an open-source package designed to easily bring machine-learning to atomistic calculations [35]**. This allows one to predict (or really, interpolate) calculations on the potential

energy surface, by first building up a regression representation from a "training set" of atomic images. The Amp calculator works by first learning from any other calculator (usually quantum mechanical calculations) that can provide energy and forces as a function of atomic coordinates. Depending upon the model choice, the predictions from Amp can take place with arbitrary accuracy, approaching that of the original calculator. Amp is designed to integrate closely with the Atomic Simulation Environment (ASE).

DeePMD-kit, a package written in Python/C++ that has been designed to minimize the effort required to build deep learning based representation of potential energy and force field and to perform molecular dynamics [36]**. DeePMD-kit is interfaced with TensorFlow (`https://www.tensorflow.org`), one of the most popular deep learning frameworks, making the training process highly automatic and efficient. On the other end, DeePMD-kit is interfaced with high-performance classical molecular dynamics and quantum (path-integral) molecular dynamics packages, i.e., LAMMPS and the i-PI, respectively. Thus, upon training, the potential energy and force field models can be used to perform efficient molecular simulations for different purposes.

Jpred (`http://www.compbio.dundee.ac.uk/jpred4`)[37] predicts secondary structural elements for an amino acid sequence using machine learning approaches, based on known 3D structural information as available in the Protein Data Bank. Machine learning approaches can predict protein-ligand binding accurately [38]**.

4. Visualization/Editing software

4.1. Micromolecules

GausView. GaussView is a commercial graphical interface used with Gaussian to build molecules or reactive systems. GaussView incorporates an excellent molecule builder which enables even very large molecules to be rapidly sketched in and then examined in three dimensions. You can also optionally add hydrogens automatically to structures originating from PDB files with excellent

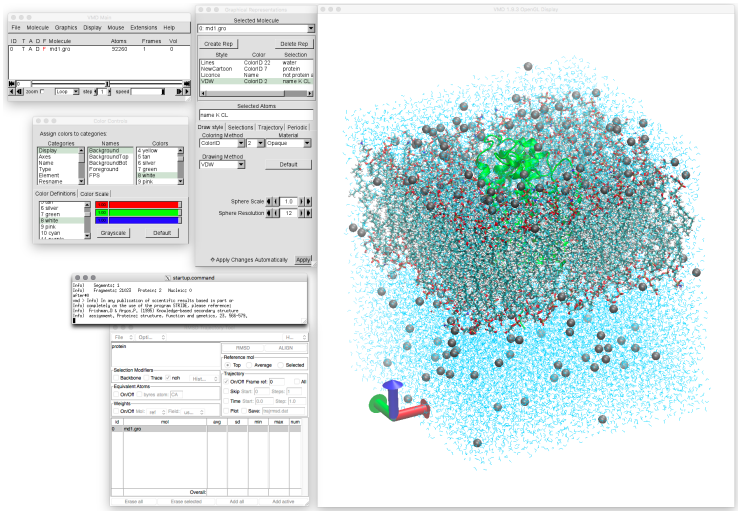

Figure 1: Screenshot of VMD on macOS showing a protein embedded in a membrane with water and ions at both sides.

reliability. The calculation is specified by pointing and clicking to build the molecule, and using pull-down menus to select the calculation type, level of theory and basis set. It aids in the creation of Gaussian input files, enables the user to run Gaussian calculations from a graphical interface without the need for using a command line instruction, and helps in the interpretation of Gaussian output (e.g., you can use it to plot properties, animate vibrations, visualize computed spectra, etc.).

ADF. The commercial Amsterdam Density Functional (ADF) software package is used by both industrial and academic researchers worldwide in computational quantum chemistry. The ADF-GUI modules include ADFview to display 3D (volume) data such as electron densities, orbitals and electrostatic potentials, ADFspectra to show spectra calculated by ADF like IR and excitation spectra, ADFMovie to follow geometry steps of geometry optimizations, IRC calculations, and ADFdos to show density-of-states graphs.

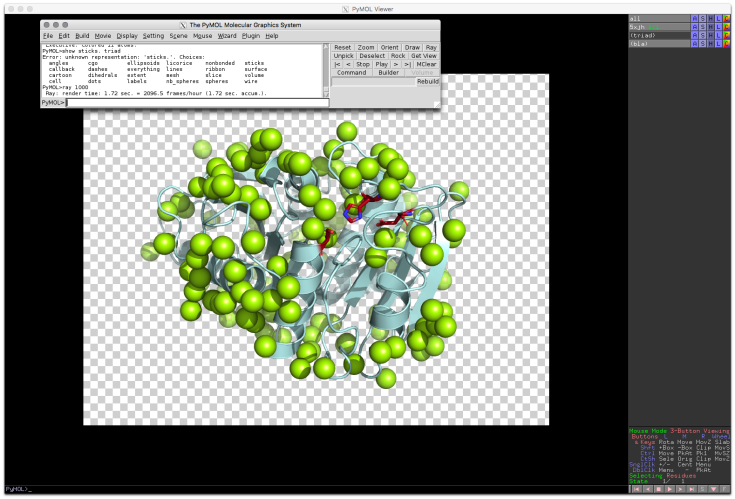

Figure 2: Screenshot of PyMOL on macOS showing a crystal structure of an enzyme with the active site highlighted and surrounded by crystalline water molecules.

4.2. Macro-molecules

VMD. VMD is designed for viewing and analyzing molecular dynamics data of biological systems such as proteins, nucleic acids, lipid bilayer assemblies, etc [39]. It also includes tools for working with volumetric data and sequence data. The functionality can be easily extended using python and Tcl scripts as VMD includes embedded Tcl and Python interpreters. Figure 1 shows a screen shot of VMD while visualizing an MD simulation of a protein embedded in a membrane.

PyMOL. PyMOL [40] is an open source molecular visualization system for (bio)molecular systems [39]. The software can produce high-quality 3D images of micromolecules and biological macromolecules by reading in structural models and volumetric data such as electron density maps. The software can easily be extended with python based scripts provided by users. Figure 2 shows a screen shot of PyMOL while rendering a high resolution image of an enzyme.

Chimera. Chimera is a program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular

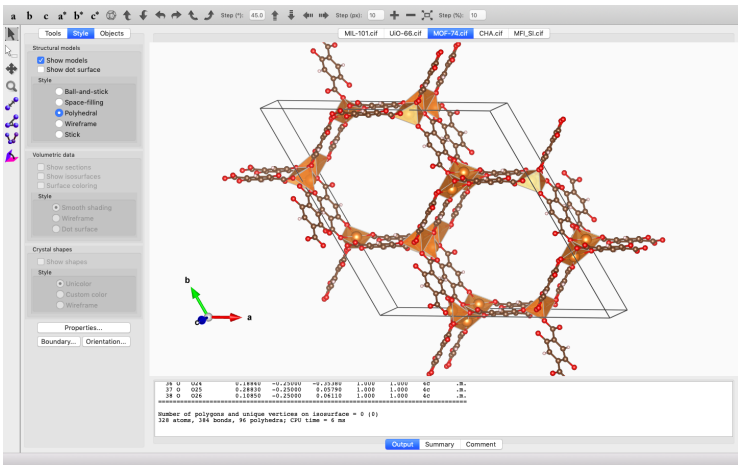

Figure 3: Screenshot of VESTA 3 on macOS showing the unit cell of a MOF-74 metal-organic framework.

assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. High-quality images and movies can be created. [41]

Webviewers. NGL Viewer is a web application for molecular visualization, aiming to display biological macromolecules [42]. 3Dmol.js is a modern JavaScript library for visualizing molecular data (`http://3dmol.csb.pitt.edu/`). This light-weight (macro)molecular visualization tool integrates easily into webpages and in particular Jupyter Notebooks. Notebook integration can be done with py3Dmol (see `https://pypi.org/project/py3Dmol/`).

4.3. Material science

VESTA. VESTA is a 3D visualization program for structural models, volumetric data such as electron/nuclear densities, and crystal morphologies [43]*. VESTA can deal with multiple structural models, volumetric data, and crystal morphologies in the same window. Figure 3 shows a screenshot of the program. It supports lattice transformation from conventional to non-conventional lattice by using matrix transformations (also used to create super- and sublattices). Transparent isosurfaces can overlap with structural models and isosurfaces can be colored on the basis of another physical quantity.

Encifer. enCIFer enables users to validate CIFs and ensure their files are format-compliant for deposition with journals and databases or for storage in laboratory archives [44]. It can visualise structure(s) in the CIF, including displacement ellipsoids, perform distance, angle or torsion measurements, and features symmetry-equivalence colouring.

jMol. Jmol is a free, open source molecule viewer for students, educators, and researchers in chemistry, biochemistry, physics, and materials science [45]. The JmolApplet is a web browser applet that can be integrated into web pages. It is ideal for development of web-based courseware and web-accessible chemical databases. The Jmol application is a standalone Java application that runs on the desktop. The JmolViewer can be integrated as a component into other Java applications. jMOL has support for unit cell and symmetry operations.

Avogadro. Avogadro is an advanced molecule editor and visualizer designed for cross-platform use in computational chemistry, molecular modeling, bioinformatics, materials science, and related areas [46]*. Avogadro features include Open Babel import of chemical files, input generation for multiple computational chemistry packages, crystallography, and biomolecules.

Crystalmaker. CrystalMaker is a commercial package that can build any kind of crystal or molecular structure quickly and easily [47]. It can visualize volumetric data from 3ED, CASTEP, Gaussian CUBE, DEN, GRD, GULP, VASP, Voxel, XSF files. CrystalMaker lets you transform the unit cell, changing the lattice type, building a supercell, moving the origin, or applying an arbitrary matrix transformation.

iRASPA. iRASPA is a visualization package (with editing capabilities) aimed at material science [48]*. Figure 4 shows a screenshot of the program. iRASPA supports crystallographic operations like space group detection and finding the primitive cell, and extensively utilizes GPU computing. For example, void-fractions and surface areas can be computed in a fraction of a second for small/medium structures and in a few seconds for very large unit cells. It can

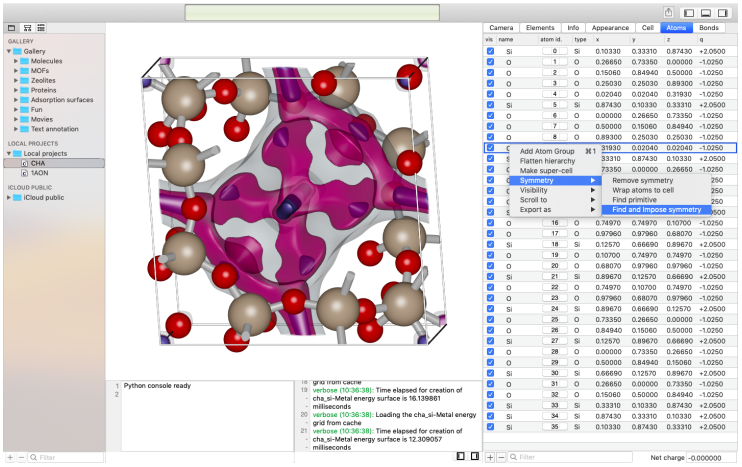

Figure 4: Screenshot of iRASPA 1.1.12 on macOS showing the primitive unit cell of a CHA-type zeolite with three adsorption surfaces showing the shape of the cavity, the diffusion paths, and the adsorption sites.

handle large structures (hundreds of thousands of atoms), including ambient occlusion, with high frame rates.

4.4. Simulation suites

Chemistry Unified Language Interface (CULGI). CULGI offers a professional modeling software package, in combination with extensive service and contract research. The software (available for Windows and Linux) covers all aspects of multiscale modeling in chemistry. It ranges from quantum chemistry to coarse-grained modeling and from chemical informatics to thermodynamics. Figure 5 shows a screenshot of the program. A feature of CULGI software is the concept of scripted workflows. Workflows can be edited through either their proprietary graphical scripting editor or Python scripting.

Software for Chemistry and Materials (SCM). The Amsterdam Modeling Suite (AMS) commercial package offers DFT, semi-empirical, reactive force fields and fluid thermodynamics all with an integrated GUI, a powerful AMS driver and python scripting tool PLAMS. AMS is particularly popular for studying complicated research questions in catalysis, spectroscopy, (bio)inorganic chemistry,

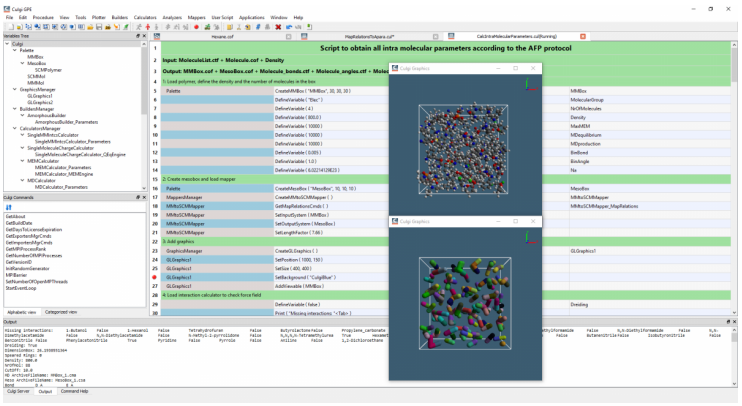

Figure 5: Screenshot of the CULGI scripting interface in the Graphical Programming Environment.

heavy element chemistry, surface science, nanoscience and materials science in general. AMS contains the following compute engines: the ADF DFT code applicable to many areas of chemistry and materials science, especially spectroscopy and inorganic chemistry, a periodic DFT code called BAND, fast approximate methods like DFTB and MOPAC to study large molecules and big periodic systems, bond order based ReaxFF to study reaction dynamics in large complex systems, and COSMO-RS to predict thermodynamic properties of solutions and mixtures.

SCIENOMICS. SCIENOMICS is a software and services company specialized in materials modeling and simulations. It offers building, visualizing, and analysis tools in one user interface: (a) Materials and Process Simulation (MAPS) for building realistic models of all types of materials, (b) SIMULATE accesses world-leading simulation engines and (c) ANALYZE for key properties to predict and screen materials behavior under different conditions. The excellent builders within the MAPS platform provide graphical interfaces for model building of any type of materials and contains a sketcher, and builders for crystals, carbon nanotubes, surfaces, interfaces, (cross-linked) polymers, amorphous materials, and meso-scale particles, lamellas or layers.

Materials Studio. Materials Studio (MS) is a commercial modeling package and simulation environment designed to allow researchers in materials science and chemistry to predict and understand the relationships of a material's atomic and molecular structure with its properties and behavior [49]. It is developed and distributed by BIOVIA (formerly Accelrys). Modeling and simulation methods in MS are: Quantum mechanics (DMol3, Castep, Gaussian), atomistic modeling QM/MM (QMERA) and MD (Discover, GULP, Forcite plus), mesoscale modeling (MesoDYN, DPD, Mesocite), crystal modeling (Reflex, Reflex Plus, Reflex QPA, X-Cell), correlations methods (QSAR, Synthia). It also includes a sorption module, a job management system, different builders for polymers, crystals, surfaces, and nanostructures, and crystallographic tools for space group detection and supercells.

Gabedit. Gabedit is a graphical user interface to computational chemistry packages like Gamess-US, Gaussian, Molcas, Molpro, MPQC, OpenMopac, Orca, PCGamess and Q-Chem It can display a variety of calculation results including support for most major molecular file formats. The advanced "Molecule Builder" allows to rapidly sketch in molecules and examine them in 3D. Graphics can be exported to various formats, including animations.

Winmostar. Winmostar is a commercial structure modeler and visualizer for chemistry simulations. Modeling and simulation methods in Winmostar are: Quantum mechanics interface to GAMESS/Firefly, NWCHem, Gaussian, SMASH and Pair Interaction Orbital analysis (PIO), MD interfaces to Gromacs, LAMMPS, and Amber, solid state physics of solids with interfaces to Quantum ESPRESSO, OpenMC, and FDMNES, and semi-emperical quantum chemistry via a MOPAC interface. It includes a job management system, and molecule-, polymer-, nanocluster- and slab-builders,

ChemAxon. ChemAxon develops chemical and biological software that provides solutions for the biotechnology and pharmaceutical industries. Core capabilities are structure visualization and management, property prediction, virtual syn-

thesis, screening and drug design. Products, like Marvin (a desktop chemical editor), are licensed free of charge for academic use.

Maestro. Maestro is a versatile modeling environment for use in pharmaceutical, biotechnology, and materials science research by Schrödinger and includes e.g. PyMOL and Quantum ESPRESSO.

Cosmologic. Cosmologic develops a set of COSOMO-related utilities, like COSMOtherm which implements COSMOS-RS (a quantum chemistry based equilibrium thermodynamics method to compute thermodynamic properties of fluids), COSMObase (high quality collections of pre-calculated compound information needed for COSMO-RS calculations), COSMOconf (a flexible tool box for conformer generation), COSMOsim3D (for automatic and unsupervised field-based ligand-ligand alignment), and TURBOMOLE (fast ab initio electronic structure calculation software that provides integration with COSMO-RS).

MedeA Software. Materials Design, Inc. develops atomistic simulation software and services for materials, includes a comprehensive graphical user interface to set up, run and analyze multi-step VASP calculations, GIBBS a forcefield-based Monte Carlo code for the prediction of fluid properties, and modules for LAMMPS, Gaussian, and MOPAC. It provides a comprehensive set of builders, including interface and amorphous material builders.

5. Tools for manuscript preparation

In general, there are two types of software to write a manuscript, Microsoft Office (or similar software like Open Office and Pages) and latex. Both have advantages and disadvantages when working with multiple authors on the same manuscript. Google Docs and Apple's Pages allows multiple users to work on the same document at the same time. When sharing files in Dropbox, Microsoft software offers sharing utilities. When using latex, sharing documents and allowing multiple users can be done via Github or via Overleaf. The latter is web-based software designed with latex documents in mind and with several

authors working on it at the same time. Opening a document in overleaf shows the source code as well as the compiled manuscript. Like GitHub, Overleaf facilitates version control.

Article searching is facilitated by citation index databases like Web of Science, Scopus, JSTOR, ScienceDirect, and Google Scholar [50]. Common reference management programs are RefWorks and EndNote, and the freely available Zotero, Mendeley and CiteULike. Bookends is a full-featured bibliography, reference, and information management system for macOS. The stored references can include attachments like the article pdf and supporting information data.

6. Operating systems: macOS, Windows, Linux

Dual-booting is a way to have several operating systems installed next to eachother. Linux is now also natively available on 64-bits windows 10 (minimum version is the Anniversary Update Version 1607) using the "Windows Subsystem for Linux". Installing an X server like Xming will allow graphical linux applications to appear on your Windows desktop. An alternative solution is to run virtualization software. Commercial options are VMWare and Parallels for macOS, and a freely available option is VirtualBox from Oracle. A downside, however, is the lack of OpenGL (>3.0) and OpenCL support. Lastly, we would like to mention that almost all open source linux software is also available on macOS using Homebrew and MacPorts.

Acknowledgements

This work was supported by the European Research Council through an ERC Starting Grant (ERC2011-StG-279520-RASPA). TJHV acknowledges NWO-CW for a VICI grant.

References

[1] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comput. Phys. 117 (1995) 1–19.

[2] P. Eastman, M. Friedrichs, J. Chodera, R. Radmer, C. Bruns, J. Ku, K. Beauchamp, T. Lane, L. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. Shirts, V. Pande, Openmm 4: A reusable, extensible, hardware independent library for high performance molecular simulation, J. Chem. Theory Comput. 9 (2013) 461–469.

[3] M. Abraham, T. Murtola, R. Schulz, S. Páll, J.C.Smith, B. Hess, E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX 1 (2) (2015) 19–25.

[4] S. Pirhadi, J. Sunseri, S. Koes, Open source molecular modeling, J. Mol. Graph. Model. 69 (2016) 127–143, Annotation: This work provides a topic perspective on open source molecular modeling.

[5] B. Kozlikova, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek, H.-C. Hege, Visualization of Biomolecular Structures: State of the Art Revisited, Vol. Computer Graphics Forum, Wiley Online Library, 2016.

[6] S. Hall, F. Allen, I. Brown, The crystallographic information file (CIF) - a new standard archive file for crystallography, Acta Crystallogr A 47 (1991) 655–685.

[7] R. Grosse-Kunstleve, N. Sauter, N. Moriarty, P. Adams, The computational crystallography toolbox: Crystallographic algorithms in a reusable software framework, J. Appl. Cryst. 35 (2002) 126–136.

[8] U. Shmueli, SPGGEN: a computer program for retrieval of space-group information in several settings and generator-containing space-group symbols, J. Appl. Cryst. 49 (2016) 1370–1376.

[9] A. Togo, I. Tanaka, Spglib: a software library for crystal symmetry search, https://arxiv.org/abs/1808.01590.

[10] M. Aroyo, J. Perez-Mato, D. Orobengoa, E. Tasci, G. de la Flor, A. Kirov, Crystallography online: Bilbao crystallographic server, Bulgarian Chemical Communications 43 (2) (2011) 183–197.

[11] D. Hicks, C. Oses, E. Gossett, G. Gomez, R. Taylor, C. Toher, M. Mehl, O. Levy, S. Curtarolo, AFLOW-SYM: Platform for the complete, automatic and self-consistent symmetry analysis of crystals, Acta Cryst. A74 (2018) 184–203.

[12] D. Weininger, SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36.

[13] D. Evans, History of the harvard chemdraw project, Ang. Chem. Int. Ed. 53 (42) (2014) 11140–11145.

[14] S. Kim, P. Thiessen, E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. Shoemaker, J. Wang, B. Yu, J. Zhang, S. Bryant, Pubchem substance and compound databases, Nucleic Acids Research 44 (D1) (2016) 1202–1213.

[15] D. Earl, M. Deem, Toward a database of hypothetical zeolite structure, Ind. Eng. Chem. Res. 45 (2006) 5549–5454.

[16] Y. Chung, J. Camp, M. Haranczyk, B. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. Farha, D. Sholl, R. Snurr, Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput computation of nanoporous crystals, Chem. Mater. 26 (21) (2014) 6185–6192.

[17] D. Nazarian, J. Camp, D. Sholl, A comprehensive set of high-quality point charges for simulations of metal-organic frameworks, Chem. Mat. 28 (3) (2016) 785–793.

[18] C. Baerlocher, L. McCusker, D. Olson, Atlas of zeolite framework types, 6th Edition, Elsevier Science, Amsterdam, 2007.

[19] S. Altschul, W. Gish, M. Webb, E. Myers, D. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[20] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[21] B. Webb, A. Sali., Comparative protein structure modeling using modeller, Curr. Protocols Bioinf. 54 (2016) 5.6.1–5.6.37.

[22] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. Heer, T. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, Swiss-model: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (2018) W296–W303.

[23] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nuclic Acids Res. 25 (1997) 3389–3402.

[24] B. Bienfait, P. Ertl, JSME: a free molecular editor in javascript, J. Cheminformatics 5 (2013) 24–.

[25] N. O'Boyle, M. Banck, C. James, C. Morley, T. Vandermeersch, G. Hutchison, Open babel: An open chemical toolbox, J. Chem. Inf. 33 (2011) 1–14, Annotation: Open Babel is an open, collaborative project allowing anyone to search, convert, analyze, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas.

[26] M. Rutter, C2x: A tool for visualisation and input preparation for castep and other electronic structure codes, Comp. Phys. Commun. 225 (2018) 174–179.

[27] J. Wang, R. Wolf, J. Caldwell, P. Kollman, D. Case, Development and testing of a general amber force field, J. Comput. Chem. 25 (2004) 1157–1174.

[28] Automatic atom type and bond type perception in molecular mechanical calculations, J. Mol. Graph. Mod. 25 (2006) 247260.

[29] V. Lesch, D. Diddens, C. Bernardes, B. Golub, A. Dequidt, V. Zeindlhofer, M. Sega, C. Schröder, ForConX: A forcefield conversion tool based on XML, J. Comput. Chem. 38 (9) (2017) 629–638.

[30] H. Jubb, A. Higueruelo, B. O.-M. no, W. Pitt, D. Ascher, T. Blundell, Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures, J. of Molecular Biology 429 (3) (2017) 365–371.

[31] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, V. S. Pande, Mdtraj: A modern open library for the analysis of molecular dynamics trajectories, Biophysical Journal 109 (2015) 1528–1532.

[32] M. Carrillo-Tripp, L. Alvarez-Rivera, O. Lara-Ramírez, F. Becerra-Toledo, A. Vega-Ramírez, E. Quijas-Valades, E. González-Zavala, J. González-Vázquez, J. García-Vieyra, N. Santoyo-Rivera, S. Chapa-Vergara, A. Meneses-Viveros, HTMoL: Full-stack solution for remote access, visualization, and analysis of molecular dynamics trajectory data, Journal of Computer-Aided Molecular Design 32 (8) (2018) 869–876.

[33] J. Tiemann, R. Guixà-González, P. Hildebrand, A. Rose, MDsrv: Viewing and sharing molecular dynamics simulations on the web, Nature Methods 14 (2017) 1123–1124.

[34] E. W. Lemmon, I. Bell, M. L. Huber, M. O. McLinden, NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 10.0, National Institute of Standards and Technology (2018). doi:https://dx.doi.org/10.18434/T4JS3C.
URL https://www.nist.gov/srd/refprop

[35] A. Khorshidi, A. Peterson, Amp: A modular approach to machine learning in atomistic simulations, Comp. Phys. Commun. 207 (2016) 310–324, An-

notation: Amp is an open-source package designed to easily bring machine-learning to atomistic calculations.

[36] W. Han, Z. Linfeng, H. Jiequn, E. Weinan, DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics, Comp. Phys. Commun. 228 (2018) 178–184, Annotation: DeePMD-kit is a package written in Python/C++, designed to minimize the effort required to build deep learning based model of interatomic potential energy and force field and to perform molecular dynamics.

[37] A. Drozdetskiy, C. Cole, J. Procter, G. Barton, Jpred4: a protein secondary structure prediction server, Nucleic Acids Res. 43 (2015) 389–394.

[38] L. Colwell, Statistical and machine learning approaches to predicting protein-ligand interactions, Current Opinion in Structural Biology 49 (2018) 123–128, Annotation: This reviews summarizes the current state of the art on machine learning approaches to predicting protein-ligand interactions.

[39] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, Journal of molecular graphics 14 (1) (1996) 33–38.

[40] Schrödinger, LLC, The PyMOL molecular graphics system, version 1.8.

[41] E. Pettersen, T. Goddard, C. Huang, G. Couch, D. Greenblatt, E. Meng, T. Ferrin, UCSF chimera–a visualization system for exploratory research and analysis, Journal of computational chemistry 25 (13) (2004) 1605–1612.

[42] A. Rose, A. Bradley, Y. Valasatava, J. Duarte, A. Prlić, P. Rose, NGL viewer: Web-based molecular graphics for large complexes, Bioinformatics 34 (21) (2018) 3755–3758.

[43] K. Momma, F. Izumi, Vesta 3 for three-dimensional visualization of crystal, volumetric and morphology data, Journal of Applied Crystallography 44 (6) (2011) 12721276, Annotation: VESTA is a 3D visualization program for

structural models, volumetric data such as electron/nuclear densities, and crystal morphologies.

[44] F. Allen, O. Johnson, G. Shields, B. Smith, , M. Towler, CIF applications. XV. enCIFer: a program for viewing, editing and visualizing CIFs, J. Applied Crystallographics 37 (2004) 335–338.

[45] A. Herraez, Biomolecules in the computer: Jmol to the rescue, Biochem. Mol. Biol. Educ. 34 (4) (2006) 255–261.

[46] M. Hanwell, D. Curtis, D. Lonie, T. Vandermeersch, E. Zurek, G. Hutchison, Avogadro: An advanced semantic chemical editor, visualization, and analysis platform, J. Cheminform. 4 (1) (2012) 17, Annotation: Avogadro is an advanced molecule editor and visualizer designed for cross-platform use in computational chemistry, molecular modeling, bioinformatics, materials science, and related areas. It offers flexible high quality rendering and a powerful plugin architecture.

[47] D. Palmer, M. Conley, Crystalmaker.

[48] D. Dubbeldam, S. Calero, T. Vlugt, iRASPA: GPU-accelerated visualization software for materials scientists, Mol. Simulat. 44 (8) (2018) 653–676, Annotation: This works describes a document-based visualization package that allows collaboration on a shared document and a CloudKit-based access to the CoRE MOF database.

[49] M. Meunier, Introduction to materials studio, EPJ Web of Conferences 30 (2012) 04001.

[50] P. Jasco, As we may search – comparison of major features of the web of science, scopus, and google scholar citation-based and citation-enhanced databases, Current Science 89 (9) (2005) 1537–1547.