Development of Open Data Analysis Tool for Science and Engineering (ODAT-SE)

Takeo Hoshi[1,2], Akito Nakano[1,2], Tatsumi Aoyama [3],
Yuichi Motoyama[3] and Kazuyoshi Yoshimi[3]

[1] *National Institute for Fusion Science,*
322-6 Oroshi-cho, Toki city, Gifu Prefecture, 509-5292, Japan.
[2] *The Graduate University for Advanced Studies, SOKENDAI,*
322-6 Oroshi-cho, Toki city, Gifu Prefecture, 509-5292, Japan.
[3] *The Institute for Solid State Physics, The University of Tokyo, Chiba 277-8581, Japan.*

We developed an open-source software Open Data Analysis Tool for Science and Engineering (ODAT-SE), by the PASUMS project in FY2024 as a major upgrade of 2DMAT (https://www.pasums.issp.u-tokyo.ac.jp/2dmat/). ODAT-SE (2DMAT) was developed for the data analysis of advanced experimental measurements and offers five analysis methods: (i) grid search, (ii) Nelder-Mead optimization, (iii) Bayesian optimization, (iv) replica exchange Monte Carlo method, and (v) population-annealing Monte Carlo method. The parallel computation with ODAT-SE was carried out on the ISSP supercomputers and other supercomputers such as Fugaku. The present report reviews ODAT-SE and a recent application study to determine the surface structure of the 3×3-Si phase on the Al (111) surface by positron diffraction and core-level photoemission spectroscopy.

1. Introduction

One of the fundamental topics in computational science is data analysis in experimental measurements. An open-source data-analysis framework 2DMAT [1] and related tools have been developed for the data analysis in experimental measurements of TWO-Dimensional MATerial structures by the PASUMS project of ISSP at FY2020 and FY2021 and other projects [1-8]. Two-dimensional materials have been a central issue in materials science because of the emergence of physical or chemical phenomena specific to their dimensionality, with promising applications in nanotechnology [9]. 2DMAT is a unified platform for the analysis of different experimental measurement techniques, total-reflection high-energy positron diffraction (TRHEPD) [10], surface X-ray diffraction (SXRD) [11], and low-energy electron

diffraction (LEED) [12] experiments.

Under the PASUMS project in FY2024, the software underwent a redesign, emerging with a new name, ODAT-SE (Open Data Analysis Tool for Science and Engineering) [13], and can be used as an interdisciplinary data-analysis framework beyond solid-state physics.

The present report contains an overview of ODAT-SE for the methodological background, the new features introduced at FY2024, a recent application study, a summary, and a future outlook.

2. Methodological background

The methodological background of the data analysis is an inverse problem [5], in which an observed quantity Y is modeled as a function of a target quantity X ($Y_{\mathrm{cal}} \equiv Y_{\mathrm{cal}}(X)$). Here, Y and X are assumed to be high-dimensional vectors ($X \equiv (X_1, X_2, \dots, X_n), Y \equiv (Y_1, Y_2, \dots, Y_m)$). The function $Y_{\mathrm{cal}}(X)$ is called a direct problem solver and should be prepared for each measurement technique. A typical analysis procedure is an optimization procedure of the residual difference function $F(X;Y) \equiv |Y - Y_{\mathrm{cal}}(X)|^2$ and the optimal solution is written as $X^* \equiv \mathrm{argmin}_X F(X;Y)$, as shown schematically in Fig. 1 (a). In the data analysis in experimental measurements, however, the quantities X and Y always contain uncertainty, and one should often consider the uncertainty quantitatively by Bayesian inference, in which the probability distribution $P(Y|X)$, called likelihood function, is modeled, and the posterior probability distribution $P(X|Y)$ is obtained as a high-dimensional histogram using Monte Carlo methods, as shown in Fig. 1 (b).

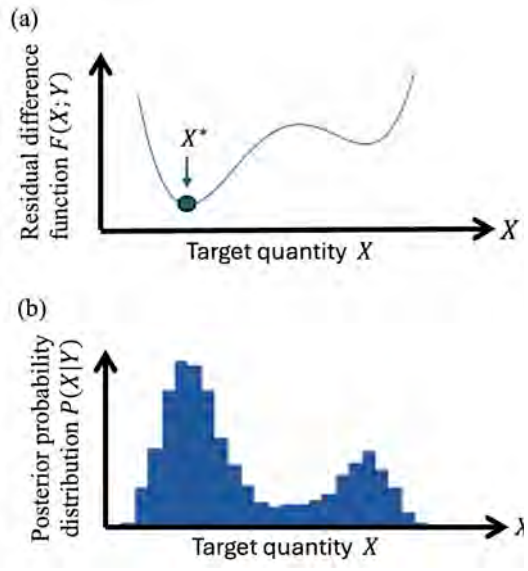

Fig. 1 Schematic diagrams of (a) residual difference function $F(X;Y)$ in optimization procedure and (b) posterior probability distribution $P(X|Y)$ in Bayesian inference

ODAT-SE offers the following five analysis methods: (i) Grid search method: The grid-search method is an algorithm for searching for the minimum value of $F(X;Y)$ by computing $F(X;Y)$ for all candidate points in the parameter space, X prepared in advance. In ODAT-SE, the set of candidate points is equally divided and automatically assigned to each process for trivial parallel computation. (ii) The Nelder-Mead optimization method [5,14]: The method, also known as the

downhill simplex method, is a gradient-free optimization method in which the gradient $\nabla_X F$ is not required. The method is reduced to an iterative local update, and a proper initial guess is needed to avoid local minimum trapping. (iii) Bayesian optimization method: The method gives a black-box optimization, and ODAT-SE uses a Bayesian optimization library, PHYSBO [15,16]. See references [15,16] for details. (iv) Replica exchange Monte Carlo (REMC) method [17] and (v) population-annealing Monte Carlo (PAMC) method [18]: The two methods, commonly parallelized Monte Carlo methods and gives the posterior probability distribution $P(X|Y)$, as in Fig. 1(b), through Bayes' Theorem $P(X|Y) = P(Y|X)P(X)/P(Y)$. The PAMC method is suitable for massively parallel computation with 10^4 MPI processes or more [5].

In ODAT-SE, the four methods, except the Nelder-Mead method, are implemented by parallel computation, which is efficient not only for personal computers but also for supercomputers. So far, the numerical computations have been carried out not only on the supercomputers at ISSP but also on the Wisteria-Odyssey supercomputer at the Information Technology Center, University of Tokyo under the Interdisciplinary Computational Science Program in the Center for Computational Sciences, University of Tsukuba, and the Fugaku supercomputer under the HPCI projects (hp210228, hp210267, hp220248, hp230304, hp240304).

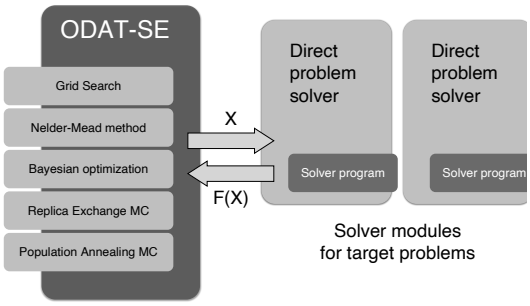

Fig. 2 Schematic diagram of the code structure of ODAT-SE

3. New features of ODAT-SE

This section explains the new features of ODAT-SE. The code structure of ODAT-SE was drawn schematically in Fig. 2. The original architecture in 2DMAT was tightly coupled with specific experimental techniques, limiting its flexibility and reusability across other scientific fields. In ODAT-SE, the architecture explicitly separates "direct problems" (i.e., the physical or statistical models representing the system under investigation) from the optimization or search algorithms. This modular approach enables researchers to apply ODAT-SE not only to 2D material analysis but also to diverse fields; users can easily add their own direct problem solvers or search algorithms tailored to their research needs.

To ensure continuity from 2DMAT, the direct problem solvers originally developed for 2DMAT have been preserved and organized in a dedicated 2DMAT GitHub

repository [19]. These models can still be used within the ODAT-SE framework with minimal modification—users simply need to install them as add-on components. Furthermore, a wide range of sample programs and usage examples has been located at the ODAT-SE Gallery [20], maintained by the ISSP Data Repository. This gallery provides ready-to-use scripts and templates, making it easier for users to adopt ODAT-SE for their specific scientific problems.

One of the significant changes in the transition from 2DMAT to ODAT-SE was the license shift from GPL v.3 to MPL v.2.0 [21]. While the GPL ensured strong copyleft—requiring any derived work to be released under the same license—it also posed barriers to integration with other software or for commercial use. In contrast, the MPL adopts a "file-level copyleft" model, allowing proprietary use and redistribution as long as modifications to MPL-licensed files are disclosed. This change has made ODAT-SE more attractive and accessible to both academic and industrial users, facilitating broader adoption and smoother integration into diverse software ecosystems.

4. A recent application study for the determination of the surface structure of the 3×3-Si phase on the Al (111) surface

This section is devoted to a review of a recent application study for the determination of the surface structure of the 3×3-Si phase on the Al (111) surface by the multiple usages of TRHEPD and core-level photoemission spectroscopy [22]. TRHEPD [10] is a novel experimental probe for two-dimensional materials and has been actively developed in the last decade at large-scale experimental facilities at the Slow Positron Facility (SPF), Institute of Materials Structure Science (IMSS), High-Energy Accelerator Research Organization (KEK) [23]. Fig. 3 shows a schematic diagram of the TRHEPD method. In this method, the observed data Y comes from the diffraction pattern seen on the screen, and the target quantity X is the position of surface and subsurface atoms at a depth of less than one nanometer.

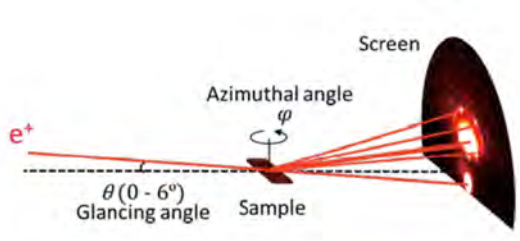

Fig. 3 Schematic diagram of the TRHEPD experiment [23]

ODAT-SE (2DMAT) was used for the data analysis of the TRHEPD experiment under the one-beam condition, in which the specular (00) spot intensities were dominated by the surface-normal component, z-component, of atomic positions. Fig. 4 shows the flat silicene structure model, a candidate of the structure.

In the model, the 3×3 surface unit cell contains eight Si atoms of the surface layer, labeled by '1', '2', …, '8' in Fig. 4, and nine Al atoms of the first sub-surface layer. The residual difference $F(X;Y)$ is given by the difference between the observed and calculated diffraction data. The PAMC analysis was carried out with the seventeen independent parameters of the z coordinates of all the eight atoms (z_1, z_2,.., z_8) of the surface Si layer and all the nine atoms of the first Al layer. As a result, the optimal structure was obtained as the flat structure for both the surface Si layer and the first subsurface Al layer, as shown in Fig. 4. See the original paper [12] for details.

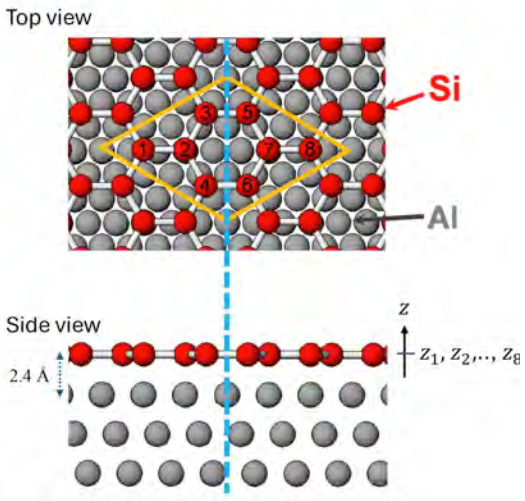

Fig. 4 Top and side views of the schematic diagram of the flat silicene structure model for the 3×3-Si phase on the Al (111) surface. A 3×3 unit cell is depicted. [22]

The analysis indicates the importance of a global search, such as PAMC, because the residual difference function $F(X;Y)$ has many local minima. The existence of local minima is demonstrated in Fig. 5, in which the grid search is carried out only with the surface Si atoms under the constraint of $z_1 = z_2 = z_3 = z_4 = z_5 = z_6 = z_8$ and the residual difference function F is drawn as a contour plot as the function of the common value of $z_1, z_2, z_3, z_4, z_5, z_6, z_8$ and the value z_7. One can find that the global minimum is located at $z_1 = z_2 = z_3 = z_4 = z_5 = z_6 = z_7 = z_8 =$ 2.4 Å, which corresponds to the flat surface structure of Fig. 4. One can also find that a local minim is located at $z_1 = z_2 = z_3 = z_4 = z_5 = z_6 = z_8 = 2.4$ Å and $z_7 = 3.6$ Å, which corresponds to a bucked surface structure. If one uses an iterative local-update algorithm, like the Nelder-Mead algorithm, the optimization algorithm may fail to find the global minimum because of the trapping on local minima.

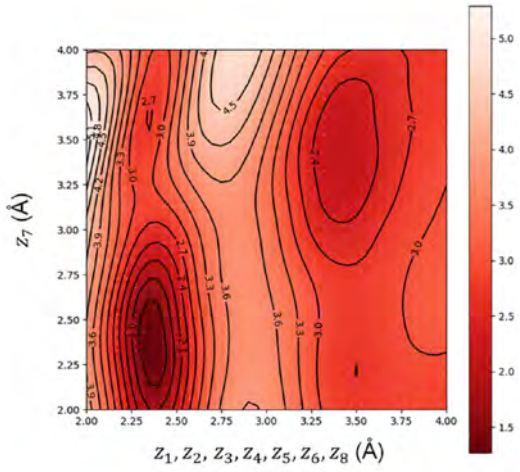

Fig. 5 Contour plot of the residual difference function in TRHEPD for the 3×3-Si phase on the Al (111) surface [22]

Here, we discuss the crucial role of the multiple usage of TRHEPD and core-level photoemission spectroscopy. In the surface analysis by TRHEPD, it should be noted that Al and Si are located next to each other in the periodic table and, thus, have similar scattering potentials for a positron beam. The observed TRHEPD data at the Si surface layer appeared nearly identical without or with Al substitution. This indicated that the surface structure model created uncertainty regarding embedded Al atoms. Si core-level photoemission spectroscopy was also carried out to overcome the above uncertainty because the Si atoms could be bonded with neighboring Si and/or Al atoms. This creates different chemical environments for these Si sites that should result in corresponding chemical shifts in Si core-level photoemission spectra.

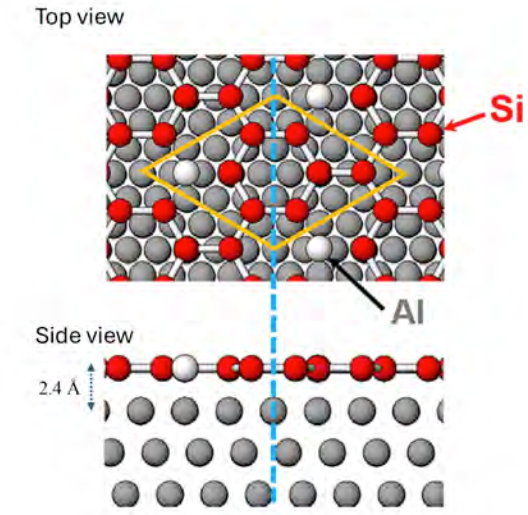

Fig. 6 Top and side views of the schematic diagram of the Al-embedded silicene model structure determined by the multiple usage of TRHEPD and core-level photoemission spectroscopy experiments. [22]

The analysis with TRHEPD and core-level photoemission spectroscopy is concluded in the Al-embedded silicene model structure in Fig.6, in which the eight sites in the 3 × 3 unit cell of the surface layer were occupied by one Al atom and seven Si atoms.

5. Summary and Future Outlook

An open-source data-analysis framework, Open-Data Analysis Tool for Science and Engineering (ODAT-SE), was developed as a major upgrade of 2DMAT. The framework offers various algorithms, mainly for the optimization and Bayesian inference as a global search. ODAT-SE has been redesigned to provide better flexibility and reusability across other scientific fields. A recent application study is reviewed for the determination of surface structure by TRHEPD and core-level photoemission spectroscopy. The analysis indicates the importance of global search and the multiple usage of TRHEPD and core-level photoemission spectroscopy.

A future outlook is the interdisciplinary use of ODAT-SE beyond solid-state physics. Very recently, the present authors launched a new project, the Moonshot R&D Program [24], in the field of fusion science, and ODAT-SE will be used in the project.

References

[1] https://www.pasums.issp.u-tokyo.ac.jp/2dmat/

[2] Kazuyuki Tanaka, Takeo Hoshi, Izumi Mochizuki, Takashi Hanada, Ayahiko Ichimiya, Toshio Hyodo, Development of data-analysis software for total-reflection high-energy positron diffraction (TRHEPD), Acta Physica Polonica A, 137, 188 (2020).

[3] Takeo Hoshi, Daishiro Sakata, Shotaro Oie, Izumi Mochizuki, Satoru Tanaka, Toshio Hyodo, Koji Hukushima, Data-driven sensitivity analysis in surface structure determination using total-reflection high-energy positron diffraction (TRHEPD), Computer Physics Communications 271, 108186/1-7 (2022).

[4] Takashi Hanada, Yuichi Motoyama, Kazuyoshi Yoshimi, Takeo Hoshi, sim-trhepd-rheed – Open-source simulator of total-reflection high-energy positron diffraction (TRHEPD) and reflection high-energy electron diffraction (RHEED), Computer Physics Communications 277, 108371/1-10 (2022).

[5] Yuichi Motoyama, Kazuyoshi Yoshimi, Izumi Mochizuki, Harumichi Iwamoto, Hayato Ichinose, Takeo Hoshi, Data-analysis software framework 2DMAT and its application to experimental measurements for two-dimensional material structures, Computer Physics Communications 280, 108465/1-11 (2022).

[6] Hisashi Kohashi, Harumichi Iwamoto, Takeshi Fukaya, Yusaku Yamamoto, Takeo Hoshi, Performance prediction of massively parallel computation by Bayesian inference, JSIAM Letters 14, 13-16 (2022).

[7] Kazuyuki Tanaka, Izumi Mochizuki, Takashi Hanada, Ayahiko Ichimiya, Toshio Hyodo, Takeo Hoshi, Two-stage data-analysis method for total-reflection high-energy positron diffraction (TRHEPD), JJAP Conf. Series 9, 011301/1-9 (2023).

[8] Shuhei Kudo, Yusaku Yamamoto, Takeo Hoshi, A fast and efficient computation method for reflective diffraction simulations, Computer Physics Communications 296, 109029/1-9 (2024).

[9] Monatomic Two-Dimensional Layers: Modern Experimental Approaches for Structure, Properties, and Industrial Use, edited by Iwao Matsuda, Elsevier (2018).

[10] Yuki Fukaya, Atsuo Kawasuso, Ayahiko Ichimiya, and Toshio Hyodo, Topical Review: Total-reflection high-energy positron, diffraction (TRHEPD) for structure determination of the topmost and immediate sub-surface atomic layers, Journal of Physics D: Applied Physics 52, 013002/1-19 (2019)

[11] R. Feidenhans'l, Surface structure determination by X-ray diffraction, Surface Science Reports 10, 105-188 (1989).

[12] M. A. V. Hove, W. Moritz, H. Over, P. Rous, A. Wander, A. Barbieri, N. Materser, U. Starke, G.A. Somorjai, Automated determination of complex surface structures by LEED, Surface Science Reports 19, 191-229 (1993).

[13] https://github.com/issp-center-dev/ODAT-SE
[14] J. A. Nelder and R. Mead, A Simplex Method for Function Minimization, The Computer Journal 7, 308–313 (1965).
[15] https://www.pasums.issp.u-tokyo.ac.jp/physbo/
[16] Yuichi Motoyama, Ryo Tamura, Kazuyoshi Yoshimi, Kei Terayamam, Tsuyoshi Ueno, Koji Tsuda, Bayesian optimization package: PHYSBO, Computer Physics Communications 278, 108405/1-8 (2022).
[17] Koji Hukushima and Koji Nemoto, Exchange Monte Carlo Method and Application to Spin Glass Simulations, Journal of the Physical Society of Japan 65 1604-1608 (1996).
[18] Koji Hukushima, Yukito Iba, Population annealing and its application to a spin glass, AIP Conference Proceedings 690, 200-206 (2003),
[19] https://github.com/2DMAT
[20] https://isspns-gitlab.issp.u-tokyo.ac.jp/takeohoshi/odat-se-gallery
[21] https://www.mozilla.org/en-US/MPL/2.0/
[22] Yusuke Sato, Yuki Fukaya, Akito Nakano, Takeo Hoshi, Chi-Cheng Lee, Kazuyoshi Yoshimi, Taisuke Ozaki, Takeru Nakashima, Yasunobu Ando, Hiroaki Aoyama, Tadashi Abukawa, Yuki Tsujikawa, Masafumi Horio, Masahito Niibe, Fumio Komori, Iwao Matsuda, Surface structure of the 3×3-Si phase on Al(111), studied by the multiple usages of positron diffraction and core-level photoemission spectroscopy, Physical Review Materials 9, 014002/1-11 (2025).
[23] https://www2.kek.jp/imss/spf/eng/
[24] https://ms10ds.nifs.ac.jp/