User-friendly workflows in quantum chemistry

Sonja Herres-Pawlis,* Alexander Hoffmann
Fakultät für Chemie und Pharmazie, Department Chemie
Ludwig-Maximilians-Universität München
Butenandtstr. 5-13, 81377 München, Germany
sonja.herres-pawlis@cup.uni-muenchen.de

Akos Balasko, Peter Kacsuk
MTA SZTAKI, Computer and Automation Research Institute
Hungarian Academy of Sciences
Budapest, Hungary

Georg Birkenheuer
Paderborn Center for Parallel Computing
University of Paderborn
Warburger Str. 100, 33098 Paderborn, Germany

Andre Brinkmann
Zentrum für Datenverarbeitung
University of Mainz
Saarstraße 21, 55122 Mainz, Germany

Sandra Gesing, Jens Krüger
Center for Bioinformatics Tübingen
University of Tübingen
Sand 14, 72076 Tübingen, Germany

Richard Grunzke
Zentrum für Informationsdienste und Hochleistungsrechnen
Technische Universität Dresden
Zellescher Weg 12-14, 01062 Dresden, Germany

Lars Packschies
Regionales Rechenzentrum
Universität zu Köln
Weyertal 121, 50931 Germany

Gabor Terstyansky, Noam Weingarten
Centre for Parallel Computing, School of Electronics and Computer Science
University of Westminster
115 New Cavendish Street, London W1W 6UW, *UK.*

***Abstract*—** Quantum chemical workflows can be built up within the science gateway MoSGrid (Molecular Simulation Grid). Complex workflows required by the endusers are dissected into smaller workflows which can be combined freely to larger meta-workflows. General quantum chemical workflows are described here as well as the real use case of a spectroscopic analysis resulting in an enduser desired meta-workflow. All workflow features are implemented via WS-PGRADE and submitted to UNICORE. The workflows are stored in the MoSGrid repository and ported to the SHIWA repository.

***Keywords*—** Quantum chemistry, Workflows, MoSGrid, Service Grids, DCIs

I. INTRODUCTION

Molecular Simulation Grid (MoSGrid) [1-2] is a workflow-oriented Science Gateway for researchers from chemistry, biology and physics which enables the access to high-performance computing (HPC) facilities. MoSGrid targets to bring more researchers into the Grid by reducing the initial hurdle of using computational chemistry software on distributed computing infrastructures (DCIs). It provides workflow tools for chemists which facilitate their computational work.

Here, we present quantum chemical workflows with different complexity from rather fundamental workflows to rather sophisticated entities which can – in best case – be dissected into meta-workflows built up of small fundamental workflows. The use cases have been identified during enduser requirements analyses within MoSGrid.

This workflow-oriented study helps a better understanding of quantum chemical processes, the needs of chemists and the productive efficiency of WS-PGRADE [3] as used in MoSGrid and the SHIWA repository [4-6].

II. BACKGROUND: WORKFLOW-ENABLED SCIENCE GATEWAYS

Workflow-enabled science gateways deal with the problem of supporting the management of workflows in a user-friendly way. The MoSGrid science gateway has been developed on top of WS-PGRADE (Web Services Parallel Grid Runtime and Developer Environment) [3], which employs the portal framework Liferay [7] and forms the highly flexible user interface of gUSE (grid User Support Environment) [8]. The MoSGrid portal offers a graphical workflow manager. Commonly used simple and complex workflows can be stored in recipe repositories and be made available for every user. All users can develop, improve, publish and use workflows for their everyday tasks. As a result of these efforts the variety of

application cases increases, making the use of computational chemistry tools easier for less experienced users at the same time.

The MoSGrid science gateway is part of the SCI-BUS project (SCIentific gateway Based User Support) [9] and directly connected to the SHIWA project (Sharing Interoperable Workflows for large-scale scientific simulations on Avaiblabe DCIs) [4-6].

SHIWA is a European FP7 project which aims at developing workflow systems interoperability technologies. SHIWA develops a Workflows Repository enabling sharing of executable workflow artifacts among user communities. It sets up an execution environment (SHIWA Simulation Platform) to execute these workflows on various Distributed Computing Infrastructures (DCIs). It also enables the creation of meta-workflows composed of workflows from different workflow management systems and it develops techniques for workflow languages translation. It is envisaged to port workflows from MoSGrid to the SHIWA Workflows Repository.

III. TECHNICAL DETAILS

A. *UNICORE*

The UNICORE grid middleware [10] offers a complete stack of tools; a graphical user interface allows to create jobs and workflows and submit them to a UNICORE grid which can consist of several clusters. UNICORE middleware services manage jobs and authenticate and authorize users. A service running on logins nodes of clusters communicates with these to run jobs for users.

In the MoSGrid project a new submitter for UNICORE was developed and contributed to gUSE [8]. It allows the submission of workflow tasks to UNICORE grids. This way the jobs can be easily distributed to clusters all over Europe. The submitter also includes functionality to index metadata. For this the UNICORE metadata service is instructed to automatically index available metadata at the end of a workflow. This makes the metadata findable of later use.

Furthermore WS-PGRADE, as the graphical user interface to gUSE, was extended to support the UNICORE incarnation database (IDB). On the one hand users are enabled to easily select tools installed on clusters to be used in workflows. Only tools available on at least one cluster can be selected. On the other hand jobs will only be submitted to cluster where chosen tools are available. The application does not have to be transmitted to the cluster, instead already installed applications are used. The user also does not have to know where the applications are installed on a cluster or on which cluster it is installed.

The MoSGrid science gateway enables the user to easily find data again. This functionality consists of a search field where terms are entered. When a term matches metadata associated to data, this data is displayed and can be selected for further analysis.

B. *Liferay*

Portals are characterized by requiring only a computer connected to the Internet and an installed web browser on the users' side. Portal frameworks aid developers with standard features like login procedures and pre-defined portlets. The open-source portal framework Liferay implements the standard JSR168 and its successor JSR286 and thus supports the re-usability of portlets for different portal frameworks that employ the same standard (e.g., Pluto). It holds responsible for the graphical appearance and user management and can be extended by portlets tailored specifically to the users' needs. It is widely used by scientific communities applying grid and cloud infrastructures, so-called DCIs (Distributed Computing Infrastructures).

C. *gUSE and WS-PGRADE*

All portals relying on remote computational resources require some kind of framework connecting the portal instance with clusters, clouds, or grids. gUSE [8] provides a large set of high-level services for the management of workflows on DCIs enabling DCI virtualization and scalable operability for the users. This includes the workflow interpreter and workflow storage responsible for the handling of scientific workflows. The application repository holds information about executable programs, while the information system keeps track about submitted jobs and user credentials. In addition, two application programming interfaces (API) interfaces (Application Specific Module API and Remote API) are provided to create application-specific science gateways according to the needs of the user community. The connection to remote DCI's is established by the so-called DCI-Bridge. The DCI-Bridge contains submitters managing the job submission to various DCIs (e.g., Globus Toolkit, BOINC). Via the project MoSGrid a specific UNICORE submitter has been developed enabling access to UNICORE grid resources [10]. Furthermore, it offers the unique feature compared to the other submitters that the available tools on a DCI can be selected without the need for uploading executables to the DCI.

WS-PGRADE is the web portal exposed to users. It offers a graphical user interface for creating, modifying, invoking, and monitoring workflows. Different client APIs of the different gUSE services are employed to convert user requests into web service calls specific for gUSE. These communication sequences are well hidden from the users behind JSR286-compliant portlets. The MoSGrid science gateway has been extended with specific portlets for the application domains quantum chemistry, molecular dynamics, and docking. The life-cycle of a workflow is managed via the Application Specific Module API (ASM). Thus, developers are able to focus on the layout and additional features of the portlet during the development process, while the workflow management is handled internally by the gUSE services. The WS-PGRADE graphical user interface is indeed utilizable by chemists who have no informatics expertise. A graph editor offers a visual access to the design of the workflow parts (tasks, input and outputs ports, connections). The subsequent "real" workflow definition is designed in an almost intuitive way of clicking through the steps. So, the step from the pure enduser to being a workflow developer becomes facilitated. At the moment, the endusers are expected to import already

developed workflows in their domain within MoSGrid or just use those which are fixed in the domain-specific portlet. It is planned to allow the MoSGrid users to develop workflows themselves and share them with the whole community.

D. Workflow Utilization

The MoSGrid science gateway aims on mapping complex chemical recipes to computer simulations that orchestrate molecular simulation codes. Chemical recipes contain several steps, beginning with the mixture of chemicals, followed by the cooking process, and the analysis of the results. During an analysis, the chemical reaction is observed and steered. After finishing one reaction, the resulting substance might be used as starting point for one or several further recipes.

MoSGrid provides two approaches to create and use workflows via the science gateway.

One approach is realized by domain specific portlets, available for molecular dynamics, quantum chemistry, and protein-ligand docking. They are hiding the chemical simulation codes, workflows, and IT infrastructures. The graphical workflow editor of WS-PGRADE is the implementation of the second approach. It allows the conception of complex and specialized workflows. Independent on the creation process, the workflows can be submitted to various distributed compute infrastructures (DCIs).

To map the recipes MoSGrid's workflows consist of several atomic tasks. The user has to select or create the workflows, upload and adapt the input, invoke the simulations on the Grid or Cloud environments, monitor the workflows, and analyze the results.

The creation process covers the definition of the workflows. When a predefined workflow is selected in one of the portlets, the chemical structure of interest and its describing properties have to be defined.

A user can also adapt or build a new workflow with the WS-PGRADE-workflow editor. A data repository contains existing workflows, workflow graphs, workflow templates, and sophisticated workflow applications that can be reused for this purpose.

Following to the creation process MoSGrids science gateway checks the user input for consistency and provide a set of default parameters (e.g., the expected wall time, number of required nodes), which can be overwritten. Additionally, users can provide command line parameters for fine-grained settings.

Because different input molecule description languages are describing the simulations, e. g. the molecular structure, application, method, temperature or basis sets, different simulation codes have to be used for the simulations. The user has two alternatives for selecting the codes. While MoSGrids portlets use available simulation codes on the infrastructure, the WS-PGRADE portal also allows uploading of executables.

The gUSE workflow engine manages a submitted workflow. It can divide complex workflows to their sub-jobs and submit them to different the DCIs. For this purpose specialized submitters are connected to the grid and cloud infrastructures, desktop grids, and web services. A running simulation is observed by WS-PGRADEs monitoring system that allows viewing intermediate results and steering of the workflow by, for example an abortion mechanism to stop unpromising simulations.

IV. General Quantum-chemical Workflows

The workflows can consist of a number of sub-jobs. Several kinds of generally interesting workflows structures were identified (Figure 1). Three general workflows are:

- Parallel executed workflows, where the same computational analysis is simulated with different substances, named **high-throughput**. Experimental structural data comes from single crystal X-ray analysis in form of *ins* file. After conversion to *mol* files they can be read by numerous programmes. A further converter combines them with blank input files for the desired quantum chemical code. Until now, the conversion is done manually and a workflow would help to save time.
- Serial executions where the result of one or more simulations is the input of the next dependent workflow step, named **TS analysis**. A transition state (TS) is found in the first calculation. A converter takes the TS geometry and generates input files for frequency calculations and intrinsic reaction coordinate scans (IRC), both calculations are required to identify the true nature of the TS:
- **Parameter sweeps** use the same substance and the same reaction with an array of different parameters. Highly practical is the workflow-approach when targeting a scan of a potential energy surface (PES) where two bond distances are simultaneously scanned for analysis of this hypersurface. The functional/basis set combination is in this case always the same and is combined by the converter with the generated coordinates.

As simulation codes, Gaussian09 [11] and NWChem [12] are used. After the end of a simulation the MoSGrid science gateway allows to extract application independent information. They can be viewed in meaningful charts or graphs in the portlets. The portal also allows downloading the results and additionally storing them in the MoSGrid data repository.

V. Concrete Use Case

A highly interesting use case is the so-called spectroscopic analysis. After a first geometry optimization of the desired molecule several further simulations are performed which serve for a spectroscopic analysis. Chemists describe this in a rather complex workflow (Figure 2).

When dissecting the single steps, we identified that there are smaller workflows of fundamental quality (such as the optimization workflow) which are embedded in the larger entity. The dissection followed the principle of identifying small tasks which can be reused within other workflows. Hence, one can define several small workflows as part of a larger meta-workflow as depicted in Figure 3.

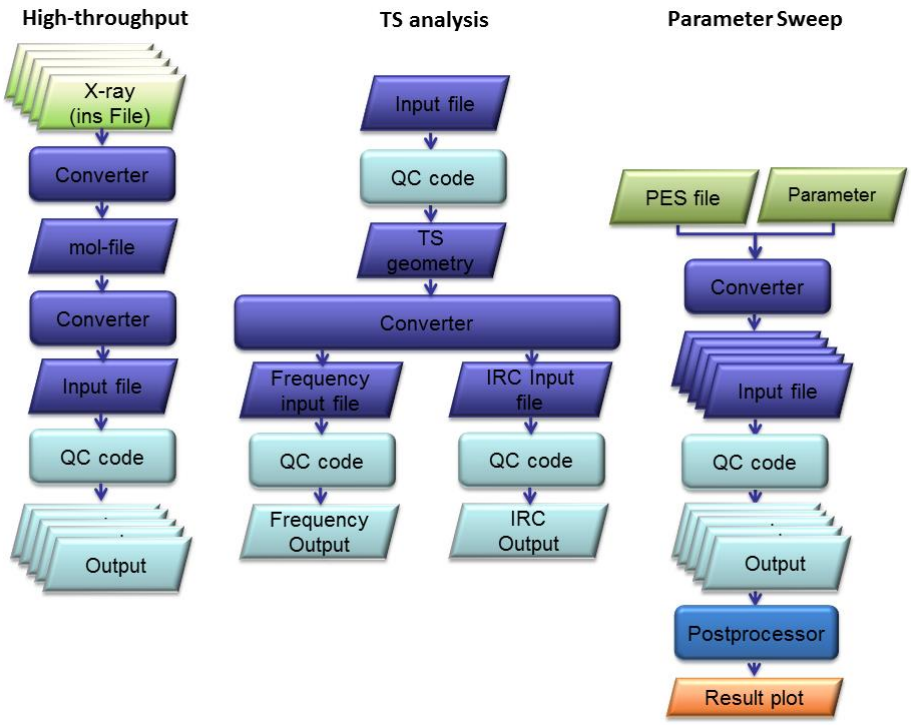

Fig. 1. General quantum chemical workflows

The workflow dissection provides with the insight that the first workflow is a simple geometry optimization (opt WF). Such a basic workflow can be reused in many more applications. The subsequent workflows are similar to each other: a converter script extracts the output geometry from the optimization output and combines it with blank input files (i.e. just lacking input coordinates) with the corresponding keywords for frequency calculations, time-dependent DFT (giving UV/Vis spectra), population analyses and subsequent calculations in solvents. All this small workflows in Figure 3 are highly valuable since they can be reused in larger quantum chemical workflows. The whole systems gains flexibility as the small workflows can be freely combined to new meta-workflows.

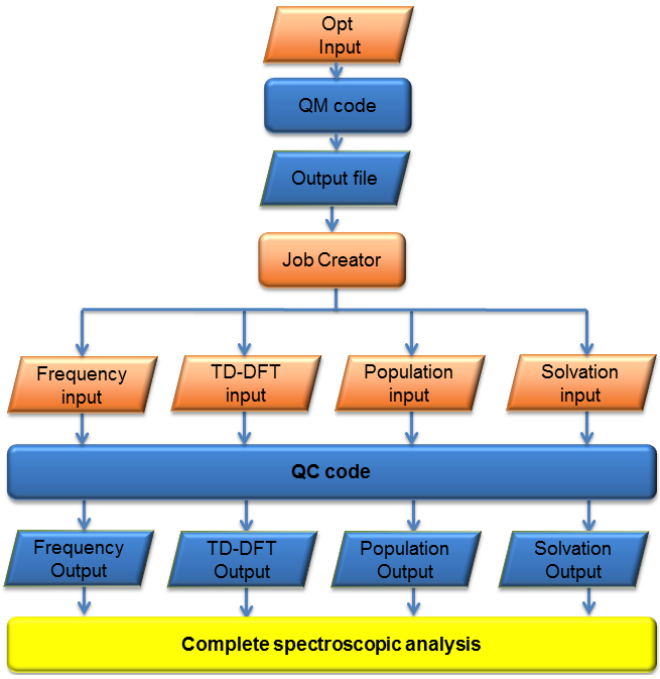

Fig. 2. Spectroscopic analysis after chemical requirement analysis

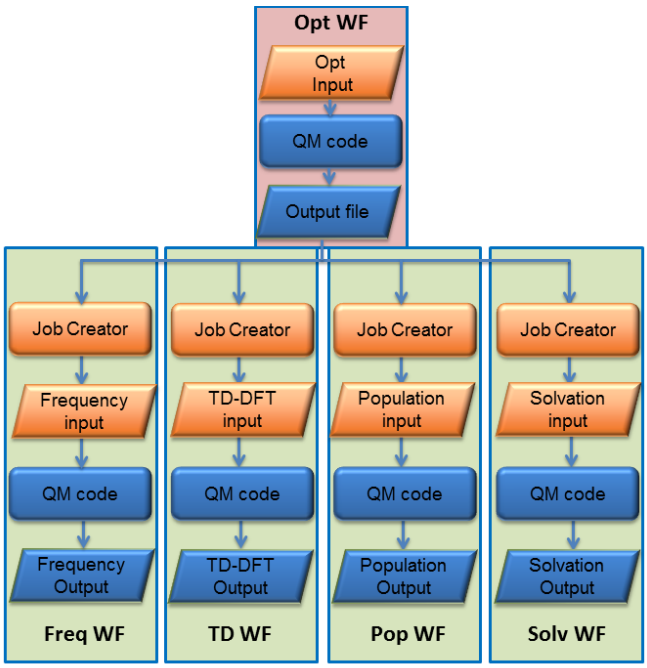

Fig. 3. Spectroscopic workflow after dissection into fundamental workflows

VI. OUTLOOK

Workflows have been proven to be a valuable tool in computational chemistry. Here, we could show that these are very practical for quantum chemical questions. We plan to dissect further complex quantum chemical workflows to smaller workflows in order to reuse the identified fundamental workflows which will be ported into the MoSGrid repository and the SHIWA repository. Then, the users can freely combine them after their requirements. The sustainability of MoSGrid and its workflows is ensured by further projects which are based on the MoSGrid initiative. Former partners of the BMBF supported MoSGrid project are now participating in SCI-BUS and ER-flow and further project proposals are under review. In fact, running MoSGrid needs only small maintenance support by the local representatives of the MoSGrid partner institutions. During the actually running projects, further advancements will be implemented in MoSGrid such as more simulation codes and a supplement in the workflow portlet enabling the free combination of workflows.

ACKNOWLEDGMENT

The authors would like to thank the BMBF (German Federal Ministry of Education and Research) for the opportunity to do research in the MoSGrid project (reference 01IG09006). The research leading to these results has also partially been supported by the European Commission's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 312579 (ER-flow) and no 283481 (SCI-BUS) and by the LSDMA project of the Helmholtz Association of German Research Centres. Special thanks are due to NGI-DE for managing the GermanGrid infrastructure.

REFERENCES

[1] S. Gesing, R. Grunzke, A. Balasko, G. Birkenheuer, D. Blunk, S. Breuers, A. Brinkmann, G. Fels, S. Herres-Pawlis, P. Kacsuk, M. Kozlovszky, J. Krüger, L. Packschies, P. Schäfer, B. Schuller, J. Schuster, T. Steinke, A. SzikszayFabri, M. Wewior, R. Müller-Pfefferkorn and O. Kohlbacher: Granular Security for a Science Gateway in Structural Bioinformatics, *Proceedings of 3rd Workshop IWSG-Life 2011*, London, UK, June 8-10, 2011, CEUR Workshop Proceedings, ISSN 1613-0073, online CEUR-WS.org/Vol-819/paper8.pdf.

[2] S. Gesing, S. Herres-Pawlis, G. Birkenheuer, A. Brinkmann, R. Grunzke, P. Kacsuk, O. Kohlbacher, M. Kozlovszky, J. Krüger, R. Müller-Pfefferkorn, P. Schäfer, T. Steinke, The MoSGrid Community – From National to International Scale. In: EGI Community Forum 2012, Munich, Germany, March 2012.

[3] Z. Farkas, P. Kacsuk, P-GRADE Portal: a generic workflow system to support user communities. Future Generation Computer Systems 27(5), 454-465 (2011).

[4] SHIWA Portal: https://shiwa-portal2.cpc.wmin.ac.uk/liferay-portal-6.1.0/de

[5] SHIWA Workflow: http://www.shiwa-workflow.eu/

[6] SHIWA Worklow Repository: http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/

[7] Inc. Liferay: Liferay. http://www.liferay.com

[8] MTA SZTAKI: gUSE. http://www.guse.hu/

[9] https://www.sci-bus.eu/

[10] A. Streit, P.Bala, A. Beck-Ratzka, K. Benedyczak, S. Bergmann, R. Breu, J.M. Daivandy, B. Demuth, A. Eifer, A. Giesler, B. Hagemeier, S. Holl, N. Lamla, D. Mallmann, A.S. Memon, M.S. Memon, M. Rambadt, M. Riedel, M. Romberg, B. Schuller, T. Schlauch, A. Schreiber, T.Soddemann, W. Ziegler, UNICORE 6 - Recent and Future Advancements. JUEL-4319 (February 2010). http://hdl.handle.net/2128/3695 (2010).

[11] M.J. Frisch, et al.: Gaussian 03, Revision C.02, (2004). Gaussian, Inc., Wallingford CT.

[12] M. Valiev, E.J. Bylaska, N. Govind, K. Kowalski, T.P. Straatsma, H.J.J. van Dam, D. Wang, J. Nieplocha, E. Apra, T.L. Windus, W.A. de Jong, "NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations" Comput. Phys. Commun. 181, 1477 (2010).