DELTA-Topology: A Science Gateway for Experimental and Computational Chemical Data Analysis using Topological Models

Sudhakar Pamidighantam
Eric Coulter
Marcus Christie
pamidigs2iu.edu
jecoulte@iu.edu
machrist@iu.edu
Cyberinfrastructure Integration Research Center,
Pervasive Technologies Institute, Indiana University
Bloomington, Indiana, USA

Aurora Clark
auclark@wsu.edu
Washington State University
Pullman, Washington, USA

ABSTRACT

Chemical data are diverse and complex, spanning point cloud data and manifolds, and occurring with potentially large dimensions. They are obtained from experimental and computational modeling, and may encode complex correlations of particle/molecular configurations and dynamic motion. It is a significant challenge to identify such correlations, reduce dimensionality, and identify the shapes and topologies of both point cloud data and chemistry-derived surfaces (e.g., energy landscapes of chemical transformation). Chemical graph theory and computational topology offer powerful new tools for the chemistry community, however, dissemination and implementation of the tools' associated algorithms and methods has been hampered by a lack of supporting infrastructure. In this manuscript, we describe the DELTA Science Gateway, which integrates several types of mathematical and topological analysis software for chemical data analysis. The focus is on energy landscape data derived from experimental and computational modeling techniques in order to understand the principles involved in structure and function. The DELTA gateway is hosted under the SciGaP project at Indiana University and is powered by Apache Airavata gateway middleware framework. The gateway provides an integrated infrastructure for simulations and analysis on XSEDE resources, as well as interactive access through a VNC client and a JupyterHub deployed on the Jetstream cloud using virtual clusters.

CCS CONCEPTS

• **Applied computing;Chemistry;Energy Landscapes; Topology;Accessibility systems and tools;Computing methodologies;Topological Data Analysis; Data Visualization**;

PEARC '21, July 18–22, 2021, Boston, MA, USA

ACM ISBN 978-1-4503-8292-2/21/07…$15.00
https://doi.org/10.1145/3437359.3465609

KEYWORDS

Science gateway, Dimensionality reduction, Interactive visualization, Harnessing Chemical Data

ACM Reference Format:
Sudhakar Pamidighantam, Eric Coulter, Marcus Christie, and Aurora Clark. 2021. DELTA-Topology: A Science Gateway for Experimental and Computational Chemical Data Analysis using Topological Models. In *Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465609

1 INTRODUCTION

Chemical systems can be modeled using fundamental physics that describe quantum mechanical electron and nuclear correlations. Classical statistical mechanical representations provide information in multiple scales and sample particle configurations; these tools can be used to predict thermodynamic and kinetic quantities and their time-dependent derivatives. Such simulation tools allow chemists to understand and corroborate experimental observations, to predict specific outcomes during reactions, and to provide a basis for the design of molecules and materials with specific physicochemical properties. From simulations, one obtains many types of data, most notably point cloud data (x,y,z coordinates of particles), as well as manifolds or surfaces such as the energy landscape (EL). Characterizing the "shape" of these data and patterns therein requires new mathematical tools and algorithms. Descriptors of Energy Landscapes Using Topological Analyses (DELTA) is an NSF-funded Harnessing the Data Revolution (HDR) Institute Frameworks project [1] that aims to advance topological data analysis (TDA) and machine learning algorithms for the study of intensive and complex data sets from chemistry, including methods that encode and learn about ELs. The ELs are 3N-dimensional surfaces that dictate the outcomes of chemical transformations, and that quickly become complicated to depict, analyze and use in predictions for large numbers of particles. Creating and disseminating novel software that combines chemical, geometrical and topological analysis in describing and predicting EL features will accelerate decision making in chemical research. As the number of dimensions relevant for a chemical transformation could be much smaller than the total dimensional space, techniques in dimensionality-reduction

such as Principal Component Analysis (PCA) [18], Uniform Manifold Approximation and Projection (UMAP) [22], and Correlation Analysis (CA) [17] provide ways to simplify and make the problem computable. Topological analysis techniques such as Morse theory [33], Persistent Homology [13] and Catastrophe theory[27] provide analytic features for ELs that describe chemical transformations.

The tools adopted and software developed by the community need an open, shareable environment in which to organize workflows and provide end-to-end access to the tools, data, and computational resources. The DELTA-Topology science gateway [2] is a community framework created to enable this environment; this manuscript describes the design, deployment and early adoption of DELTA-Topology. This gateway differentiates itself from other chemistry-oriented gateways such as SEAGrid (SEAGrid.org) through its focus on topological analysis. The manuscript is organized as follows: in Section 2, we describe the gateway design and implementation. Section 3 describes applications and their deployment and integration. Section 4 presents the deployment of interactive computing and visualization on a Cloud-based virtual cluster through JupyterHub and VNC. Section 5 concludes with an outlook for the gateway in the near future.

2 GATEWAY DESIGN AND DEPLOYMENT

The DELTA-Topology gateway is deployed using the Apache Airavata [25] middleware framework and provides reproducible and shareable job management services. It runs on SciGaP [26] gateway hosting services at Indiana University, and development support is provided through a consultation from Science Gateways Communities Institute (SGCI) [32]. An XSEDE allocation[30] provides computational gateway hosting resources. The high performance computer Stampede2 [28] provides high-end computational resources, and the Jetstream Cloud [29] provides gateway and data hosting facilities. A virtual elastic cluster on Jetstream cloud is also provisioned for on-demand and interactive workloads. Extended Collaborative Support Services (ECSS) [31] from XSEDE provided additional assistance in integrating the resources into the DELTA-Topology gateway. The Delta-Topology gateway authenticates users via the OpenID Connect (OIDC) protocol [3]. We use an instance of the Keycloak [14] identity and access management service as the OIDC provider. The Delta-Topology tenant in Keycloak is configured to integrate CILogon [11] as a federated identity provider while also allowing users to create local user accounts. We have 55 active users supporting 10 participating PI research groups.

The gateway administrator controls access to resources by granting access to user groups, with different groups having different scopes of access. Users are authorized via Airavata Sharing services [23]. The members of the "Admin" group have additional privileges and are responsible for deploying applications, creating user interfaces, monitoring and trouble shooting, and updating notifications. They are also authorized to control metadata for accessing XSEDE HPC resources through the gateway "community login," managing users; they can also troubleshoot any issues relating to user services. Gateway administration is handled through an Admin Dashboard, which provides interfaces to functions to set security credentials, define compute resource profiles, control access to resources through group-based user organization, and monitor experimental statistics. Once the gateway administrator provides approval, new users are added to a default group ("Gateway Users") that gives access to gateway resources and applications. Approved users can then use gateway services to create, monitor, share and clone experiments (computational simulations). Users can also add their own compute resource allocations in SciGaP registered resources using the "User Settings" profile functions. Users and administrators can create groups and share/provide controlled access to resources, applications, and data. For example, most students get a "Developer" role so they can deploy and test applications in the gateway. Such features make Apache Airavata a desirable middleware framework. The DELTA webinars and tutorials are hosted on the DELTA-Topology gateway for quick reference and broad outreach.

The Airavata Django Portal [15] front-end web framework deployed for the DELTA-Topology gateway integrates the Wagtail content management system (CMS) [9] to drive the theme and allow gateway managers to edit, review and preview the look and feel, save drafts, and publish pages for user consumption. The Delta-JupyterHub link on the navigation bar becomes available upon successful authentication. The Django framework gives a way to authenticate and access applications, and to submit and monitor jobs through separate dashboards. The gateway is enhanced by access to a 10TB disk volume for data hosting, and the data volume is integrated as a Globus [10] end point for use with Globus clients to upload data easily. Currently, the gateway supports 54 members of the DELTA project team and collaborators that have run close to 460 experiments so far.

3 APPLICATIONS, DEPLOYMENT, INTEGRATION AND USE

As an NSF-funded HDR Framework (conceptualization) award, the DELTA framework research teams are rapidly adopting and integrating diverse analysis software, and the software tends to evolve rapidly, as well. The basic software, such as Chemnetworks [24], is reasonably stable with few releases, but other basic analysis software in the topology is still evolving.

Application	Function	Resource Deployed on
Gudhi	Persistent Topology	Stampede2
Ripser	Vietoris–Rips persistence barcodes	Stampede2
Javaplex	persistent homology	Stampede2
Dionysus	persistent homology	Stampede2
AnalyzeTrajectory	MD Trajectory Visualization	Stampede2
Umapper	Reduced network map	Stampede2
Pathreducer	PCA based reduction	Stampede2
ChemNetworks	Networks in chemical structures	Stampede2
TTK	Topological analysis	Jetstream elastic virtual cluster
Paraview	Visualization	Jetstream elastic virtual cluster
VisIt	Visualization	Jetstream elastic virtual cluster
Scikit-TDA	Python Library	Stampede2
PagerankAnalysis	Collective Variable	Stampede2
Matlab	Multipurpose	Stampede2
VNC-run	Interactive visualization	Jetstream elastic virtual cluster
Jupyter hub	Jupyter notebooks	Jetstream elastic virtual cluster

Table 1: Applications deployed in DELTA-Topology gateway

The applications currently deployed in the DELTA-topology gateway are provided in Table 1, with their functionality and the resources that deploy them. Core application frameworks such as UMAP [22], RIPSER [12], Gudhi [20], Javaplex [8] have been installed on XSEDE's Stampede2 [28] system, and analysis applications are developed on top of the core software. Several applications

are deployed as Python libraries and could be easily called from other Python codes and as part of Jupyter notebooks. As codes are developed and committed, containers are created and deployed on a container hub when appropriate; they can subsequently be used in a singularity environment on Stampede2 when jobs are launched from the gateway. Figure 1 shows the user interface deployed for a Delta-Persistence application and an output persistent bar graph that results from a run on Stampede2 to illustrate gateway use.

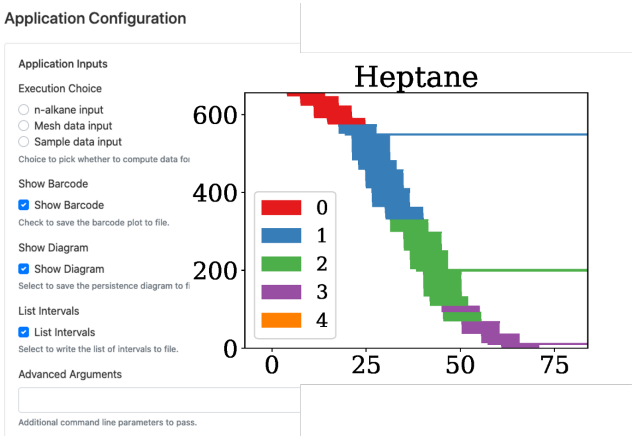

Figure 1: User interface Delta-persistence application and sub-level-set persistent bar codes based on carbon chain length of 7 for Heptane.

VNC applications start a TigerVNC [4] server on a virtual cluster on a cloud resource and send back a link to launch a local client to connect to it. This is generic but will be coupled with visualization software such as Topology tool kit (TTK) [5], VisIt [6] and Paraview [19]. Users provide parameters for the applications through user interfaces and set up computational experiments using Experiment Editor panels. Experiments (jobs) are launched on specified resources with scheduler-level job specifications such as nodes/cores, memory and wall time. The experiments are monitored through an Experiment Summary panel on launch, which is refreshed to update the status automatically. The results are also reported in the Experiment Summary panel as rendered images, text or links to outputs. The files are retrieved into a storage device dedicated to the gateway.

4 INTERACTIVE COMPUTING AND VISUALIZATION

As discussed earlier, the DELTA-Topology gateway features the ability to run interactive sessions for visualization needs through a VNC Client. This is modeled on the feature offered by TACC through their visualization portal, but operates in an elastic virtual cluster on the Jetstream cloud [29]. For this type of job, the Slurm scheduler creates a visualization node on the virtual cluster with the necessary GUI tools installed, and provides a link to the user that allows them to access the new VNC session securely through a websocket tunnel. The VNC session gives users access to a full XFCE4 desktop environment [7] with access to all computational tools installed on the cluster. The interactive node also mounts the same filesystems as any other job on the resource, allowing users access to previously-created data for visual analysis. The ParaView and VISIT tools are currently available on the cluster. The host and port ID for the server are presented in a job-specific link protected by the VNC password set by the user. This is presented to the user in the Experiment Summary panel as the ViewVNC link as shown in the left panel in Figure 2 as soon as it is generated on the server. The right panel in Figure 2 presents the view of the VNC client with a Paraview application of a model surface.

In addition to interactive jobs run through the VNC client, users can access a JupyterHub instance on the same Jetstream cloud-based elastic cluster. This resource is configured so that a new Jupyter server is launched as a Slurm job, which creates a new, dedicated node for the Jupyter session. JupyterHub is configured to authenticate users with the same Keycloak tenant as the Delta-Topology gateway. There is a short wait when the JupyterHub is setup; it is served once ready.

5 OUTLOOK

In the near future, many new features will become available in Apache Airavata and will be deployed in the DELTA-topology gateway in an automated scheme. The interactive features developed for DELTA will be generalized for the Airavata framework shortly for broader impact. Work to provide an embedded VNC client within the Django gateway front-end is underway. The user and administrator dashboards will be enhanced with search and retrieve functions for arbitrary experiments, applications, and time periods to improve data organization. The Django framework provides a well-defined extension mechanism (Django apps) that we will use to develop specific applications that require workflows with dependent inputs from other steps as in directed acyclic graphs.

ACKNOWLEDGMENTS

This work is supported by a HDR Framework grant from the National Science Foundation under grant #1934725. We would like to acknowledge XSEDE [30] for providing computing, hosting services, and extended collaborative support services. Developer support was provided by the Science Gateways Community Institute through NSF Award #1547611. SciGaP hosting services [16]and the Apache Airavata gateway middleware framework [21] were supported by NSF Award #1339774. We acknowledge Eroma Abeysinghe (IU) for assistance in the early deployment of the gateway and Sonja Gerard (WSU) for theme development and integration. We are thankful to all the DELTA senior personnel and students who contributed applications and interfaces, and who tested the deployments.

REFERENCES

[1] 2019. https://doi.org/awardsearch/showAward?AWD_ID=1934725&HistoricalAwards=false
[2] 2021. delta-topology.org
[3] 2021. , OpenID Connect pages. https://openid.net/connect/
[4] 2021. https://tigervnc.org/
[5] 2021. https://topology-tool-kit.github.io/index.html
[6] 2021. https://wci.llnl.gov/simulation/computer-codes/visit
[7] 2021. https://www.xfce.org/
[8] Henry Adams, Andrew Tausz, and Mikael Vejdemo-Johansson. 2014. javaPlex: A Research Software Package for Persistent (Co)Homology.. In *ICMS (Lecture*

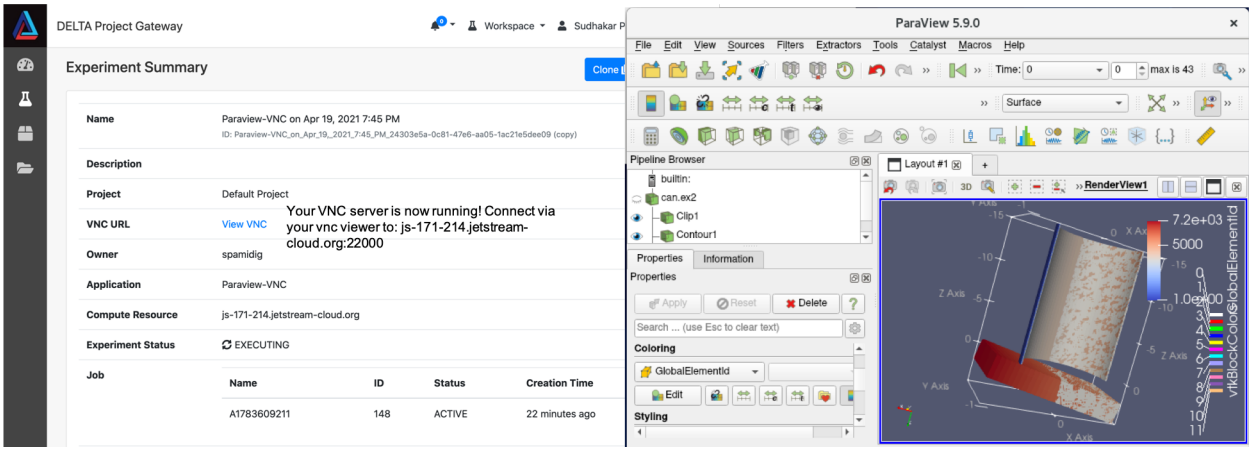

Figure 2: A VNC Job on DELTA-Topology gateway executing remotely on a Jetstream Cloud-based virtual cluster through Slurm scheduler. The "Experiment Summary" dashboard (left) provides a link that contains the VNC access information. A Paraview application launched remotely is accessed interactively (right).

Notes in Computer Science, Vol. 8592), Hoon Hong and Chee Yap (Eds.). Springer, 129–136. http://dblp.uni-trier.de/db/conf/icms/icms2014.html#AdamsTV14

[9] Stephen Paul Adithela, Marcus Christie, Suresh Marru, and Marlon Pierce. 2018. Django Content Management System Evaluation and Integration with Apache Airavata. In *Proceedings of the Practice and Experience on Advanced Research Computing* (Pittsburgh, PA, USA) *(PEARC '18)*. Association for Computing Machinery, New York, NY, USA, Article 86, 4 pages. https://doi.org/10.1145/3219104.3229272

[10] Bryce Allen, John Bresnahan, Lisa Childers, Ian T. Foster, Gopi Kandaswamy, Rajkumar Kettimuthu, Jack Kordas, Mike Link, Stuart Martin, Karl Pickett, and Steven Tuecke. 2012. Software as a service for data scientists. *Commun. ACM* 55, 2 (2012), 81–88. http://dblp.uni-trier.de/db/journals/cacm/cacm55.html#AllenBCFKKKLMPT12

[11] Jim Basney, Heather Flanagan, Terry Fleury, Jeff Gaynor, Scott Koranda, and Benn Oshrin. 2019. CILogon: Enabling Federated Identity and Access Management for Scientific Collaborations. In *International Symposium on Grids & Clouds 2019 (ISGC2019) - Network, Security, Infrastructure & Operations*, Vol. 351. https://doi.org/10.22323/1.351.0031

[12] Ulrich Bauer. 2019. Ripser: efficient computation of Vietoris-Rips persistence barcodes. (2019). arXiv:1908.02518v1 [math.AT]

[13] GUNNAR CARLSSON. 2009. TOPOLOGY AND DATA. (2009). https://www.ams.org/journals/bull/2009-46-02/S0273-0979-09-01249-X/S0273-0979-09-01249-X.pdf

[14] Marcus Christie, Anuj Bhandar, Supun Nakandala, Suresh Marru, Eroma Abeysinghe, Sudhakar Pamidighantam, and Marlon Pierce. 2017. Using Keycloak for Gateway Authentication and Authorization. https://doi.org/10.6084/m9.figshare.5483557.v1

[15] Marcus Christie, Suresh Marru, Eroma Abeysinghe, Dimuthu Upeksha, Sudhakar Pamidighantam, Stephen Paul Adithela, Eldho Mathulla, Aarushi Bisht, Shivam Rastogi, and Marlon Pierce. 2020. An extensible Django-based web portal for Apache Airavata. In *Practice and Experience in Advanced Research Computing*. 160–167.

[16] Marlon E. Pierce *et al.* 2015. Apache Airavata: design and directions of a science gateway framework. *Concurrency and Computation: Practice and Experience* 27, 16 (2015), 4282–4291. https://doi.org/10.1002/cpe.3534

[17] Gary Horlick and Gary M. Hieftje. 1978. *Correlation Methods in Chemical Data Measurement*. Springer US, Boston, MA, 153–216. https://doi.org/10.1007/978-1-4615-6734-9_4

[18] I.T. Jolliffe. 1986. *Principal Component Analysis*. Springer Verlag.

[19] Kitware, Inc. [n.d.]. ParaViewWeb. http://www.paraview.org/Wiki/ParaViewWeb.

[20] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. 2014. The Gudhi Library: Simplicial Complexes and Persistent Homology.. In *ICMS (Lecture Notes in Computer Science, Vol. 8592)*, Hoon Hong and Chee Yap (Eds.). Springer, 167–174. http://dblp.uni-trier.de/db/conf/icms/icms2014.html#MariaBGY14

[21] Suresh Marru, Lahiru Gunathilake, Chathura Herath, et al. 2011. Apache airavata: a framework for distributed applications and computational workflows. In *Proceedings of the 2011 ACM workshop on Gateway computing environments*. 21–28.

[22] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://arxiv.org/pdf/1802.03426.pdf cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap.

[23] Supun Nakandala, Suresh Marru, Marlon Piece, Sudhakar Pamidighantam, Kenneth Yoshimoto, Terri Schwartz, Subhashini Sivagnanam, Amit Majumdar, and Mark A Miller. 2017. Apache airavata sharing service: A tool for enabling user collaboration in science gateways. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*. 1–8.

[24] Abdullah Ozkanlar and Aurora E. Clark. 2014. ChemNetworks: A complex network analysis tool for chemical systems. *J. Comput. Chem.* 35, 6 (2014), 495–505. http://dblp.uni-trier.de/db/journals/jcc/jcc35.html#OzkanlarC14

[25] Marlon. E. Pierce, Suresh Marru, Eroma Abeysinghe, Sudhakar Pamidighantam, Marcus Christie, and Dimuthu Wannipurage. 2018. Supporting Science Gateways Using Apache Airavata and SciGaP Services. In *PEARC '18: Proceedings of the Practice and Experience on Advanced Research Computing*.

[26] Marlon. E. Pierce, Suresh Marru, Eroma Abeysinghe, Sudhakar Pamidighantam, Marcus Christie, and Dimuthu Wannipurage. 2018. Supporting Science Gateways Using Apache Airavata and SciGaP Services. In *PEARC '18: Proceedings of the Practice and Experience on Advanced Research Computing*.

[27] Tim Poston, Ian Stewart, and Alan Jeffrey. 1978. *Catastrophe theory and its applications*. Pitman, London; San Francisco. https://www.worldcat.org/title/catastrophe-theory-and-its-applications/oclc/780449993&referer=brief_results

[28] Dan Stanzione, Bill Barth, Niall Gaffney, Kelly Gaither, Chris Hempel, Tommy Minyard, Susan Mehringer, Eric Wernert, H Tufo, D Panda, et al. 2017. Stampede 2: The evolution of an xsede supercomputer. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*. 1–8.

[29] Craig A Stewart, Timothy M Cockerill, Ian Foster, et al. 2015. Jetstream: a self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. 1–8.

[30] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* 16, 5 (Sept.-Oct. 2014), 62–74. https://doi.org/10.1109/MCSE.2014.80

[31] Nancy Wilkins-Diehr, Sergiu Sanielevici, Jay Alameda, John Cazes, Lonnie Crosby, Marlon Pierce, and Ralph Roskies. 2016. An overview of the XSEDE extended collaborative support program. In *High Performance Computer Applications - 6th International Conference, ISUM 2015, Revised Selected Papers (Communications in Computer and Information Science, Vol. 595)*. Springer Verlag, Germany, 3–13. https://doi.org/10.1007/978-3-319-32243-8_1

[32] Nancy Wilkins-Diehr, Michael Zentner, Marlon Pierce, Maytal Dahan, Katherine Lawrence, Linda Hayden, and Nayiri Mullinix. 2018. The Science Gateways Community Institute at Two Years. In *Proceedings of the Practice and Experience on Advanced Research Computing* (Pittsburgh, PA, USA) *(PEARC '18)*. Association for Computing Machinery, New York, NY, USA, Article 53, 8 pages. https://doi.org/10.1145/3219104.3219142

[33] Edward Witten. 1982. Supersymmetry and Morse theory. *Journal of Differential Geometry* 17, 4 (1982), 661 – 692. https://doi.org/10.4310/jdg/1214437492