Topical Group on Data Science

Summer 2020 Newsletter

Message from the Editor

Kyle Wm. Hall, GDS Newsletter Editor

Hello GDS Members!
For many of us, the last few months have passed by in a surreal fashion. Much of what we had planned for the summer seems to have vanished into thin air. And yet, upheavals can be the crucible for positive change. In this vein, GDS has seen a number of exciting developments, such as the webinar series. Now more than ever, forging digital connections is vital to realizing a vibrant community, and these newsletters are one element of the GDS digital strategy.

The summer edition of the GDS newsletter contains a number of exciting elements starting with a welcome message from the new GDS Chair, Jie Ren, along with information about on-going GDS initiatives and the upcoming 2021 meetings. Dimitri Bourilkov summarizes the GDS invited session at the 2020 April Meeting, Data Science in Physics Education, which is sure to remain an exciting and relevant area in the years to come. Emine Küçükbenli provides a nice summary of atomistic modelling software based on machine learning while Sergei Kalinin introduces Bayesian inference, and illustrates how it serves as a bridge between physics and machine learning. Jerome Delhommelle highlights some of his recent work on leveraging machine learning to capture partition functions. I hope that you enjoy this edition of the GDS newsletter.

Kind Regards,

Kyle Wm. Hall

Kyle Wm. Hall is currently a Postdoctoral Fellow with Michael L. Klein at Temple University. Kyle obtained an interdisciplinary PhD from the University of Calgary under the supervision of Drs. Peter G. Kusalik and Sheelagh Carpendale. His research lies at the intersection of molecular simulations and data science, and spans data visualization, human-computer interaction, polymer physics, and crystallization.

In this Issue

How to Connect

There are a variety of ways that you can stay up to date, and informed about GDS. Follow GDS on social media if you haven't done so already!

Email
gds@aps.org

Website
https://www.aps.org/units/gds/index.cfm

LinkedIn
https://www.linkedin.com/company/apsdatascience

Twitter
https://twitter.com/apsdatascience

Facebook
https://www.facebook.com/APSDataScience

The information and views provided herein are those of the individuals involved, and do not necessarily correspond to those of APS, GDS, or the Newsletter Editor.

Message from the Chair

Jie Ren, GDS Chair

Dear GDS members,

I am excited to serve you as the GDS chair for the 2020-2021 term. This is a belated greeting, though; I was preparing to meet with many of you in-person during the 2020 March Meeting in Denver – our first major APS conference as a newly-formed topical group. Championed by GDS's founding chair, Mohammad Soltanieh-ha, in this meeting GDS had planned a full agenda of data science-themed events, from sessions and tutorials to a fun-filled night with Google Cloud. The cancellation of the March Meeting was no doubt a disappointment; but the rising threat of COVID also taught me, more than ever, the value of community and the importance of building resilience against challenges.

The world has changed quite a bit since the March Meeting cancellation, with the pandemic, lockdowns, economic uncertainties, as well as social unrests. Despite such circumstances, GDS steadily grew in membership and gained momentum with an expanding archive of webinars, thanks to the tremendous support from our members. Through the GDS webinar series, we have invited distinguished speakers across academia, industry, and national labs to share research work at the interface of data science and physics, conducted hands-on tutorials on deep learning and autonomous experimentation, and discussed state-of-the-art tools for data science education for physicists. Your great support for these events is evident through the number of live attendees and replays, as well as the busy Q&A's, highlighting the strong need for data science discussions in the community. We are committed to continue addressing such need, by further facilitating knowledge sharing and cross-conversations on data science.

With its goal to build a strong community within APS, GDS has a lot more to do and invites everyone to participate. Below is a short list of items that we hope to work on with your help. I'd like to thank the executive committee for initiating these efforts:

- **Building a diverse & inclusive community through action.** GDS represents a culture of diversity and inclusion; we believe that such value is at the heart of vibrant and productive science. I encourage you to read *GDS's statement on diversity* (page 3) and join us in the related efforts described in this statement.
- **Strengthening COVID-related efforts among physicists through community-building.** Sponsored by the APS executives, GDS has partnered with the Topical Group on Medical Physics (GMED) to build the COVID Research and Resources Group (CRRG). Details about the CRRG can be found on page 3; we welcome your suggestions and help in any form.
- **Organizing sessions and events at the 2021 March Meeting and April Meeting.** As stated on page 4, the call for submissions for the 2021 annual meetings are open, and we look forward to your nominations, abstract submissions, etc.
- **Conducting ongoing GDS Webinar series.** Please see page 5 on how to suggest new topics for the webinars and participate in the webinar organization.
- **Nominating and electing future members of the executive committee.** Nominations for GDS executive committee members for the 2021-2022 term will open in the fall; be sure to make submissions if you or someone you know share a passion for data science.

As always, we are here to serve your needs and would love to hear your suggestions. You can reach out to us any time: gds@aps.org. To stay up to date, please also follow us on LinkedIn, Facebook, and Twitter: @APSDataScience. I am grateful for having you as part of GDS. Together, we will be a welcoming, resilient community that rises to meet challenges, and that continues to strengthen the synergy between data science and physics.

With kind regards,

Jie Ren

GDS Statement on Diversity

The GDS Executive Committee

Within GDS, we are committed to fostering a diverse, equitable, and inclusive community where all members can feel safe and welcome. We stand by the APS Diversity Statement and the more recent APS Leadership Letter Condemning Racism, and firmly believe that diversity and inclusion are at the heart of vibrant and productive science. There is great richness that comes from a plurality of thought and experience. GDS treasures diversity and seeks to connect physicists from all groups who are interested in data science applications. However, in many ways, our communities are still far from environments where all groups of people are welcomed, valued, and empowered to flourish. GDS is committed to promoting more diverse, equitable, and inclusive communities. The GDS executive committee seeks to both understand and address on-going challenges, and will take the following actions in the coming months.

- Survey our membership to better understand the barriers and challenges that currently exist across the intersection of physics and data science.
- Establish an ombudsman for diversity, equity, and inclusion to further support the people within our GDS community.
- Host a webinar focused on diversity, equity, and inclusion as part of the on-going GDS webinar series, which will include a discussion on how data science can be used to expose and address issues of diversity, equity, and inclusion.
- Play an active role in identifying and recruiting speakers from historically marginalized groups to highlight the scholarly work being conducted by the full diversity of researchers.

GDS will continue to pursue and support activities that promote diversity, equity, and inclusion. The GDS executive committee always welcomes suggestions and comments from members.

Kind Regards,

The GDS Executive Committee

COVID Research and Resources Group

To better serve the physicists involved in COVID-related work, a COVID Research and Resources Group (CRRG) has been initiated through a close collaboration between GMED (Topical Group on Medical Physics) and GDS. The CRRG has been created on the APS Engage platform, through which APS intends to better understand the community needs and provide additional resources and support accordingly. Activities by the CRRG will include both broad-based COVID research webinars, and events/discussions within focused interest groups (e.g., modeling, data analytics, imaging, and technologies). It also intends to assist with outreach activities via communications with partner organizations, societies, industry corporations, and news media.

If you are actively conducting COVID-related research or are interested in joining such efforts, please make sure to join the CRRG on APS Engage. You may log into APS Engage with your My APS credentials and opt into the COVID Research and Resources Group. Please also fill out a short questionnaire to help us organize CRRG activities.

2021 March Meeting and April Meeting Program Planning

March Meeting.
The GDS Programming Committee has finalized the 2021 sorting categories. As listed in Table 1, GDS will sponsor 16 focus topics, and 6 main sorting categories in the 2021 March Meeting. In addition, GDS also has 1 invited session slot, which can support one (1) GDS wholly-owned session or two (2) co-sponsored sessions with partner units. Contributed abstract submissions are now open, and will be open until to October 23rd.

GDS will use the ScholarOne system to facilitate session organization processes for the 2021 March Meeting. Abstract submission will be handled through this platform. A link to the ScholarOne portal is provided on the APS March Meeting 2021 website. If you encounter any issue or need help with the process, do not hesitate to reach out to us at gds@aps.org.

April Meeting.
Planning activities for the next April Meeting are also warming up, with calls for invited speakers and contributed abstracts planned in the fall. This past April, GDS successfully delivered its inaugural invited session in a first-ever virtual April Meeting thanks to the great work by the session chair, Dimitri Bourilkov. Following this effort, Dimitri will continue to serve as GDS's April Meeting Program Chair for the 2021 April Meeting. Please reach out (dimi@ufl.edu) for additional information.

Table 1. GDS Sorting Categories for the 2020 March Meeting

Category ID	Description (Co-sponsors)
23.00.00	GDS Symposium Invited Speaker (Invitation Only)
23.01.00	GDS FOCUS SESSIONS
23.01.01	Big Data in Physics (GDS, DCOMP, GSNP)
23.01.02	Deep Learning for Dynamical Systems (GDS, DCOMP)
23.01.03	Deep Learning for Spectroscopy (GDS, DCOMP)
23.01.04	AI Materials Design and Discovery (GDS, DCOMP)
23.01.05	Open Science/Open Data (GDS)
23.01.06	AI & Real-World Networks (GDS, DBIO)
23.01.07	Autonomous Systems and Control (GDS)
23.01.08	AI and Statistical/Thermal Physics (GDS, GSNP, DCOMP)
23.01.09	Visualization Techniques and Systems (GDS, DCOMP)
23.01.10	Machine Learning for Quantum Matter (DCOMP, GDS, DMP)
23.01.11	Machine Learning and Data in Polymer Physics (DPOLY, DBIO, DCOMP, GDS, FIAP)
23.01.12	Emerging Trends in Molecular Dynamics Simulations and Machine Learning (DCOMP, GDS, DSOFT, DPOLY)
23.01.13	Deep Learning for Computer Vision (GDS)
23.01.14	Machine learning for biomolecular design and simulation (DPOLY, GDS, DSOFT, DBIO, DCOMP)
23.01.15	Artificial intelligence and machine learning in medicine and biomedicine (GMED, GDS)
23.01.16	Quantum machine learning (DQI, GDS)
23.02.00	GDS STANDARD SORTING CATEGORIES
23.03.00	Data Science in Physics
23.04.00	Machine Learning
23.05.00	AI and Deep Learning
23.06.00	Big Data, Data Integration and Assimilation
23.07.00	Data Science Education
23.08.00	Autonomous Experimentation

GDS Webinar Series

Initiated after the cancellation of the 2020 March Meeting, the GDS Webinar Series has not only converted most of the cancelled GDS talks, courses and tutorials to online formats, but has also extended far beyond the original conference contents and continues to bring new, exciting talks at the intersection of data science and physics. Over the past several months, GDS has developed a mature webinar operation workflow, publicized and delivered over a dozen successful webinar events featuring over 30 invited speakers in total. A cumulative list of over 1,000 unique registrants have attended our live webinars; in addition, the GDS YouTube channel with a collection of the webinar replays now has 240+ subscribers and over 3,000 unique views. We are immensely grateful for the tremendous interest in this content, which makes us even more committed to identifying topics of interest to the community and delivering contents that serve your needs.

Suggestions of topics, speakers, or volunteering to host webinars are always very welcome. In order to serve a diverse and inclusive community of physicists, GDS especially encourages the nomination of speakers who are women, members of underrepresented minority groups, and scientists from outside the United States.

Data Science & Physics Education at the April Meeting

Dimitri Bourilkov (Associate Scientist, Department of Physics, University of Florida)

The inaugural April GDS Invited Session: Data Science in Physics Education (H06), organized together with the Forum on Education, took place at the first virtual APS Meeting in April 2020. After having secured three excellent speakers, I am happy to report that two of them were able to join this lively session with over seventy attendees. Many thanks to the APS staff for their extraordinary efforts to make all this possible at a short notice. A recording of the session is available at:

https://aps-april.onlineeventpro.freeman.com/live-stream/15336053/H06-Invited-Session-Data-Science-in-Physics-Education

The first talk on "Machine Learning in Particle Physics" was given by Dr. Sergei Gleyzer from the University of Alabama. He discussed the application of state-of-the-art data science methods, like modern deep learning networks, to detector reconstruction, particle identification, real-time event filtering, and new physics searches, as well as the significant new opportunities in the particle physics field that have been enabled by data science.

In a lively Q&A, Sergei answered questions like how to avoid bias by taking the example of a mass peak search in Higgs boson decays to two photons. The machine learning tool will try to "cheat" if we feed input features that can be used to guess the mass. Two ways to counter this are: 1) preprocess the input features, e.g. by using scale invariant variables, to fight the symptoms, and 2) *a posteriori* by applying penalties to the loss function. The former is the preferable option if it is possible.

The second talk "CERN Open Data Portal for Science and Education" was presented by Dr. Matthew Bellis from Siena College. While other fields like astronomy have practiced regular releases of their data for the community to analyze after an embargo period, particle physics experiments have been slow to do the same, given the complexity of the datasets and the analysis workflows. Starting in 2014, CERN launched the Open Data Portal hosting data from the Large Hadron Collider (LHC) experiments and the OPERA neutrino experiment. Educational examples and simplified datasets are provided alongside access to the same data and tools used by the experimental analysts. Members of the theory community have used the Compact Muon Solenoid (CMS) open data to publish in journals.

In the following Q&A, Matthew answered questions on how to use the data for publications. The very permissive Creative Commons CC0 waiver is used, so the data are truly open, just cite them properly using their DOI. One help in developing high school courses online is a workshop for school teachers: Particle Physics Playground. It provides simplified real particle physics data and Python tools running in a browser using Google Colab. There will be an exciting workshop this Fall for theorists and phenomenologists to get hands-on tutorials interacting with CMS open data. Datasets are made available also for the broad ML community to explore new ideas.

Overall, this year's inaugural Data Science in Physics Education session was a success, and I look forward to future discussion facilitating greater integration of data science in physics education.

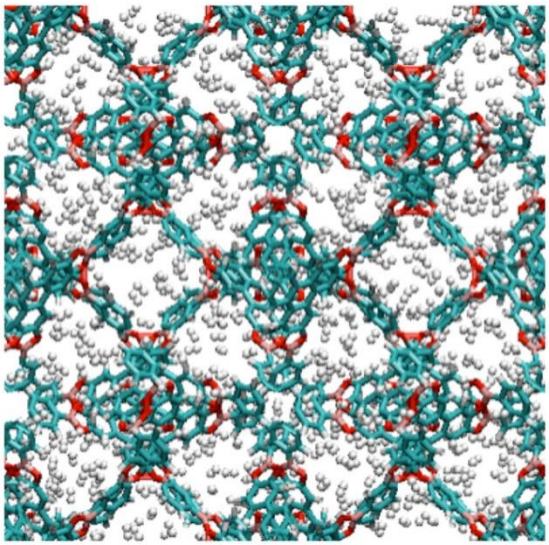
Dr. Dimitri Bourilkov was born in Sofia, Bulgaria, where he obtained his PhD in particle physics at the Institute for Nuclear Research and Nuclear Energy. He has conducted research at the largest accelerators in the world with the BIS-2experiment at Serpukhov, Russia, and the L3 and CMS experiments at LEP and LHC in Geneva, Switzerland. After starting his career in Sofia, Dr. Bourilkov has worked at the Joint Institute for Nuclear Research in Dubna, Russia, at the Radboud University in Nijmegen, the Netherlands, at ETH Zurich, Switzerland, and is now a Scientist in CMS with the University of Florida, Gainesville. He is a recipient of an Internet2 IDEA award in 2007 for developing the next generation of network-aware grids.

Machine Learned Partition Functions for the Rapid Assessment of Nanoporous Materials Properties.

Jerome Delhommelle (Associate Professor, University of North Dakota)

The partition function is a central and extremely powerful quantity in statistical mechanics, which encompasses considerable information about a system and provides access, for instance, to any of its thermodynamic properties. Recent advances in computational physics, specifically flat histogram sampling algorithms, now allow for accurate numerical calculations of the partition function. However, such sampling algorithms are extremely computationally intensive, especially for multicomponent systems, and need to be repeated for new sets of thermodynamic conditions (e.g. for every temperature). To address this challenge, we have recently introduced a diversity ensemble learning approach based on neural networks to predict the partition function of fluids adsorbed in nanoporous materials (see [1]). As detailed in our recent study, we have applied our approach to several environmental and energy applications, including carbon capture and storage in metal-organic frameworks, and hydrogen storage in covalent organic frameworks. The new insights achieved through the ensemble learning approach include rapid determination of the selectivity and the desorption free energy over a wide range of operating conditions (pressure or mole fraction of the gas feed) for each of the nanoporous materials considered. Our approach enables rapid assessment of the relative performance of nanoporous materials for storage and separation applications, as well as a measure of the energetic cost to regenerate the adsorbent.

See our paper [1] for methodological details about our data science approach and for more information about our findings.



Jerome Delhommelle is an Associate Professor in Chemistry & in Physics at the University of North Dakota. He is a former student of the Ecole Normale Superieure (Cachan, France) and has received a NSF CAREER award from the Division of Materials Research, as well as a 2012 OpenEye Young Faculty Award from the ACS (Division of Computers in Chemistry). He is also the North American Editor for the journal "Molecular Simulation", published by Taylor & Francis. Together with Dr. Desgranges, he is currently focusing on how a machine learned entropy can unravel activated processes, such as, among others, crystal nucleation

References

1. Caroline Desgranges and Jerome Delhommelle J. Phys. Chem. C, 124, 1907 (2020)
DOI: 10.1021/acs.jpcc.9b07936

.

Learning Machines in Atomistic Modelling: The Software

Emine Küçükbenli (Postdoctoral Fellow, Harvard University)

As the atomistic modelling communities within chemistry, physics and material science are embracing machine learning (ML), new methods and applications that bring ML to these fields are surging. While the communities are focused on the success and limitations of these new approaches, a new software burst is also taking place to answer the needs of the new paradigm. This perspective piece provides an overview of the current state of affairs of software development in this new and exciting research field of atomistic modelling with ML, and outlines some of the main challenges and potential developments ahead.

A diverse set of methods used in chemistry and physics for decades now fall under the umbrella term of "Machine Learning". For example, clustering is now often called unsupervised ML while kernel regression and neural networks are called supervised ML methods. Among these, even the most data- and computation-intensive approaches (i.e., neural networks) were already being tried out in the atomistic modeling literature as early as 1995 [1]. Hence the chemistry, physics and material science communities are no strangers to ML. Yet in recent years, we have seen more researchers going beyond the role of end-users, and becoming the architects behind new algorithms and software. This trend is particularly visible within computational physics in the context of two supervised learning approaches, namely modelling interatomic interactions (a.k.a. force field generation) via Gaussian Process Regression (GPR) and Neural Networks (NN).

GPR is a subclass of what is known as "kernel machines" where the kernel trick is used to learn a target function via regression. From a programming perspective, GPR is similar to other methods in this class, e.g., Kernel Ridge Regression. Implementation of kernel machines rely on operations that are common in mathematical libraries, such as inner products to build kernel matrices, solving linear equations for weights or optimization of hyperparameters. Hence, the most valuable software contributions by the physics community often come as wrappers, tools and interfaces that integrate GPR to existing modeling software. **QUIP** [2], a toolbox written in FORTRAN with a python interface, provides help building GPR-based interatomic interaction models in a format that can be used with existing molecular dynamics (MD) software. **VASP**, a proprietary first-principles electronic structure package, integrates GPR with its core code so that the data generated during ab initio calculations can be used to build ML models on the fly and accelerate MD simulations [3]. **FLARE**, an open-source software, uses the ability of GPR method to provide confidence intervals for its predictions. It detects when the ML model uncertainty crosses a threshold during a simulation, and places requests for more data, in the spirit of "active learning" [4]. Other packages that implement ML methods with similar software components to GPR also exist. For example, **MLIP** approximates potential energy surfaces via linear regression with polynomial-like functions of atomic coordinates [5]. LAMMPS, a widely popular classical MD package implements **SNAP**, another method based on linear regression that uses atomic neighbor density around a central atom as the basis of regression in contrast to other approaches [6].

An even wider landscape of software is available for neural network methods, paralleling their sophistication and complexity from a programming perspective. **RuNNer** [7] and **AENet** [8] packages, which are written in C++ and FORTRAN respectively, are perhaps the most historically relevant ones, as they were developed within research groups that were the early adopters and innovators of NN interatomic potentials. **AMP** [9] with a FORTRAN core is perhaps the first open source package to bring NN interatomic potentials to the larger community of materials modeling experts with its user-friendly interface and interoperability across many MD packages. **PROPhet** [10] (written in C++) is no longer maintained but merits a special mention; its highly efficient MD plug-in made NN potentials practical, and showcased possible ways to reduce the cost of NN potentials, which still lies around a few orders of magnitude higher than classical force fields.

The literature has seen a substantial increase in software diversity for atomistic modelling with the introduction of open source libraries written in high-level languages that are developed and maintained by

the larger ML community. For example, **PyTorch** [11], primarily developed by Facebook's AI Research, is a popular back-end for scientific applications. Its python interface allows easy scripting of network models compared to the packages written in C++ or FORTRAN highlighted earlier. **TorchANI** [12], **SchnetPack** [13] and **KLIFF** [14] are some of the PyTorch-based atomistic NN suites. They implement classical feed forward network architectures or innovatively use known network architectures for atomistic systems, such as continuous-filter convolutional NNs. Another popular backend for neural network training is **TensorFlow** [15], developed and maintained mainly by Google. Tensorflow differs from PyTorch mostly due to its dataflow programming focus. In TensorFlow, the program is modelled as a static directed acyclic graph, where operations take place as soon as their input is ready, rather than operations being stacked one after another waiting for a specifically coded order to get executed, allowing effective performance optimization options. Although static graph-based programming is not commonly encountered in scientific software packages in materials modeling, the adoption of TensorFlow appears to be wide ranging; **PANNA [**16], **DeePMD** [17], **SIMPLE-NN** [18], **ChemML** [19] are some of the interatomic potential generation packages that use TensorFlow backends.

A factor that distinguishes NN packages from one another is the input structure they are built for, similar to basis-set choices in electronic structure codes (e.g., atomic orbitals, plane-waves etc). The question of input, i.e., how to best represent an atomic system to an NN, is non-trivial. For example, an NN that learns to predict the potential energy of a periodic crystal should yield the same value, irrespective of simulation cell choice. The assumptions made in representation and mathematical description of the input constitute one of the important distinctions between the packages listed above. Recently developed libraries that implement many different descriptor types in a unified manner (e.g., **DScrib**e [20]) may enable packages to have access to several types with ease and leave room for more innovation in network architecture. Another new development towards unifying efforts is JAX-MD [21] where neural network building and MD preparation can be built-in together natively using the JAX framework which itself is built on powerful autograder technology in which native python calls can be differentiated automatically, allowing any step of an atomistic simulation to be machine-learnt without the development overhead.

As the plethora of software frameworks shows, ML in atomistic simulations has evolved from specialized in-house codes to community packages in a short time. Some important tasks still remain

Predictive Power: Many published works are still method-showcase studies of well-known systems, making the predictive power of ML models hard to estimate for research-relevant cases, especially without reports of negative results.

Cost: ML is still very computationally costly when the resources spent in data gathering, training and testing are considered, yet it is often unknown whether the gain is much more than to adding a handful more parameters to well-tested classical models.

Reproducibility: This common challenge has greater importance in ML due to difficulties with interpretability. Initiatives such as **OpenKIM** [22] that store models can help: 1) keep track of progress in the field, 2) aid with the reproducibility of new developments, and 3) could also be used to measure transferability of models between datasets.

All these challenges would benefit greatly from having a well-connected software developer community among ML researchers in atomistic modeling. The time-tested approach to building healthy communities has been workshops and conferences among developers of different codes. In light of the on-going COVID-19 pandemic, the young community dedicated to ML atomistic modeling must find virtual ways to address the issue of community building. At GDS, we are ready to facilitate such exchanges and look forward to having you joining us in such discussions.

Emine Küçükbenli is a postdoctoral fellow at Harvard John A. Paulson School of Engineering and Applied Sciences. Her research aims to model accurately and explore efficiently the vast landscape of crystal structures that atoms and molecules form. She builds numerical tools that optimize the exploration of potential energy landscapes by introducing theoretical or algorithmic improvements as well as by implementing machine learning approaches. She is a contributor and developer in open source projects such as Quantum ESPRESSO, PANNA, ESL. She is a passionate advocate for reproducibility in science.

References

1. Thomas B. Blank, Steven D. Brown, August W. Calhoun, and Douglas J. Doren; J. Chem. Phys. 103, 4129 (1995); doi: 10.1063/1.469597

2. Bartók, A. P., Csányi, G. J. Quantum Chem. 2014, 115, 1051– 1057. DOI: 10.1002/qua.24927

3. Ryosuke Jinnouchi, Ferenc Karsai, and Georg Kresse; Phys. Rev. B 100, 014105 DOI: 10.1103/PhysRevB.100.014105

4. J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky. npj Computational Materials 6, 20 (2020), DOI: 10.1038/s41524-020-0283-z

5. E.V. Podryabinkin, A. V. Shapeev, Computational Materials Science, volume 140, pages 171-180, 2017. DOI: 10.1016/j.commatsci.2017.08.031

6. A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, G.J. Tucker; Journal of Computational Physics, 285, 316-330 (2015) DOI: 10.1016/j.jcp.2014.12.018.

7. J. Behler, and M. Parrinello, Phys. Rev. Lett. 98, 146401 (2007).

8. N. Artrith and A. Urban, Comput. Mater. Sci. 114 135-150 (2016).

9. Khorshidi & Peterson, Computer Physics Communications 207:310-324, 2016. DOI:10.1016/j.cpc.2016.05.010

10. Kolb, B., Lentz, L.C. & Kolpak, A.M. Sci Rep 7, 1192 (2017). DOI: 10.1038/s41598-017-01251-z

11. A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library", in: H. Wallach, H. Larochelle,A. Beygelzimer, F. dAlch´e-Buc, E. Fox, R. Garnett (Eds.), Advances In Neural Information Processing Systems 32, Curran Associates, Inc.,2019, pp. 8024–8035.

12. Gao X, Ramezanghorbani F, Isayev O, Smith JS, Roitberg AE. (2020) DOI: 10.26434/chemrxiv.12218294.v1

13. K.T. Schütt, P. Kessel, M. Gastegger, K. Nicoli, A. Tkatchenko, K.-R. Müller. J. Chem. Theory Comput. 15, 1, 448–455 (2019) DOI:10.1021/acs.jctc.8b00908

14. M Wen, RS Elliott, EB Tadmor, (2020) URL: https://github.com/mjwen/kliff/

15. Martín Abadi et al. (2015) URL: tensorflow.org

16. R. Lot, F. Pellegrini, Y. Shaidu, E. Küçükbenli, Comp. Phys. Comm, (2020) 107402 DOI: 10.1016/j.cpc.2020.107402.

17. H. Wang, L. Zhang, J. Han, and W. E. Comp. Phys. Comm. 228 178-184 (2018) DOI: 10.1016/j.cpc.2018.03.016

18. K. Lee, D. Yoo, W. Jeong, S. Han, Comp. Phys. Comm., 242, 95-103, (2019) DOI: 10.1016/j.cpc.2019.04.014.

19. Haghighatlari, M., Vishwakarma, G., Altarawy, D., Subramanian, R., Kota, B., Sonpal, A., Setlur, S., & Hachmann, J. ChemRxiv, 8323271 (2019) DOI: 10.26434/chemrxiv.8323271.v1

20. L. Himanen, M.O.J. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, Computer Physics Communications 247, 106949 (2020) DOI: 10.1016/j.cpc.2019.106949.

21. S. S. Schoenholz and E. D. Cubuk; arXiv:1912.04232 (2019)

22. E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller and C. A. Becker, JOM, 63, 17 (2011). DOI:10.1007/s11837-011-0102-6

Bayesian Inference:
The Bridge from Physics to Machine Learning

Sergei V. Kalinin (Corporate Fellow, CNMS ORNL)

The physical sciences are underpinned by cycles of hypotheses and experiments. Based on past observations and knowledge, we seek explanations for observed phenomena. Once possible explanations and models are established, we seek to define and perform the experiments that can prove or disprove these models. However, as famously said by Lord Kelvin, "*When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science*." Hence, the question becomes how can we express knowledge in numbers?

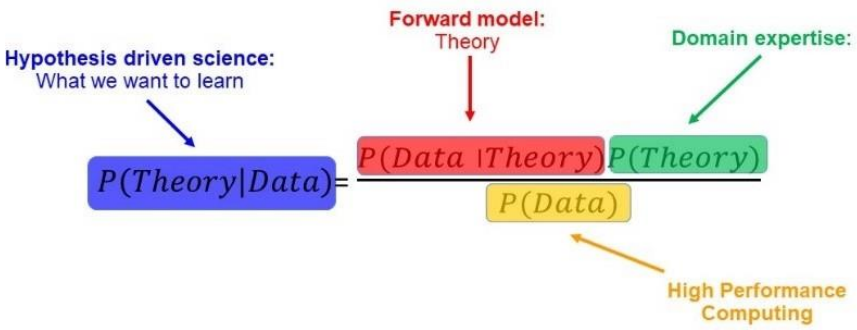
The classical approach to analyzing experimental results is grounded in frequency-based, or Fischer, statistics. Repetitions of experiments yield information on reproducibility, point estimates of parameter values, or p-values for hypotheses. However, these numbers are often difficult to define and interpret beyond simple text-book examples. In fact, the interpretation of the hypothesis as correct or incorrect can depend on an idealized experiment [1]. Unsurprisingly, classical statistics beyond elementary definitions tends to remain obscure to all but a few physicists.

Bayesian methods provide a natural framework for physicists to build a bridge from their specific domain to a broad world of statistics, and hence machine learning. Bayesian statistics relies on the concepts of prior and posterior probabilities linked via Bayes formula:

$$p(\theta_i|D) = \frac{p(D|\theta_i)p(\theta_i)}{p(D)} \qquad (1)$$

Here the notation $p(A)$ defines the probability of event A happening, and $p(A|B)$ refers to the conditional probability of A happening if B has happened. D represents the data obtained during the experiment, and the observed phenomena are assumed to be ascribable to the N possible models (hypothesis) with a set of parameters θ_i, i = 1, .., N. In this notation, $p(D|\theta_i)$ is the *likelihood* that the observed data can be generated within the model, a quantity that can be calculated directly for a known model (and with a sufficiently powerful computer). The new knowledge derived with Bayes formula is the posterior, $p(\theta_i|D)$, i.e. the probability of our model (or hypothesis) given the experimental observations. The key element of Bayesian inference is the prior $p(\theta_i)$, i.e., the probability function for the model and model parameters as known ***before*** the experiment. Finally, $p(D)$ is the denominator that defines the total space of possible outcomes. Overall, Eq. (1) signifies how the new data, D, changed our hypotheses, θ_i, about the system, directly following Lord Kelvin dictum.

While extremely powerful for classical tasks such as descriptive statistics and functional fits, Bayesian methods open fundamentally new opportunities when implemented as a part of experiment. For example, Gaussian process regression allows one to explore behaviors in functional spaces, with the coupling between the function values for different points in parameter space given by the kernel function. The regions of maximum uncertainty can be used as a basis for subsequent explorations. Combinations of



uncertainty and function optimization allow a researcher to combine exploration and exploitation in Bayesian Optimization frameworks.

Despite the intrinsic elegance of Bayes approach, its adoption by many scientific communities has been slow. The need to choose priors, $p(\theta_i)$, led to criticism of the biased nature of Bayesian methods. However, while the choice of priors is indeed heavily domain-specific, the same judgments are being made even outside Bayesian context (e.g. "are these parameter values reasonable for this material?"), albeit not in a quantitative fashion. Secondly, the physical sciences have access to vast amounts of prior knowledge in the form of known laws, past data, and observations, though such information is not necessarily in the convenient form of statistical distributions. Perhaps a good comparison here are the famous words of Naser od-Din Tusi (1201-1274), who said

Har kas ke bedanad va bedanad ke bedanad
Asb-e kherad az gombad-e gardun bejahanad
Har kas ke nadanad va bedanad ke nadanad
Langan kharak-e khish be manzel beresanad
Har kas ke nadanad va nadanad ke nadanad
Dar jahl-e morakkab'abad od-dar bemanad

Anyone who knows, and knows that he knows
Makes the steed of intelligence leap over the vault of heaven
Anyone who does not know, but knows that he does not know
Can bring his lame little donkey to the destination nonetheless
Anyone who does not know, and does not know that he does not know
Is stuck forever in the double ignorance

The everyday work of the scientists falls best under the second category, seeking the new knowledge. But there is a fourth possibility – there is much that we know but we are unaware that we know it, and Bayesian methods can open the pathway to collect and integrate that knowledge.

In the context where both Bayesian and classical methods can be used (e.g., function fitting), Bayesian methods behave no worse or better than classical least square fits, which is reassuring. However, there is another potential barrier to applications of Bayesian inference. Evaluation of the denominator in Eq. 1 requires very high dimensional integrals. Fortunately, this has become feasible for experimentally-relevant distributions in the last several years. The development of computation ecosystems in languages such as Python and R have alleviated this problem.

Where can one start the Bayesian journey? Naturally, every person has a favorite sequence of books and papers that give the necessary introduction and philosophy as well as recipes that can be applied to practical problems. One such pathway can start with an introduction to Bayesian methods by Lambert [2], and an outstanding book by Martin [3]. The latter introduces the philosophy of Bayesian methods, and gives easy to follow examples in the PyMC3 library. The book by Kruschke [1] contains an excellent comparison of frequency-based statistics and Bayesian statistics paradigms. And of course, for those who are interested in experimenting further, Python libraries such as GPTorch, Pyro, etc. open opportunities that were impossible just five years ago.

Sergei Kalinin is a corporate fellow at the Center for Nanophase Materials Sciences at Oak Ridge National Laboratory. He received his MS degree from Moscow State University in 1998 and Ph.D. from the University of Pennsylvania (with Dawn Bonnell) in 2002. His research presently focuses on the applications of big data and artificial intelligence methods in atomically resolved imaging by scanning transmission electron microscopy and scanning probes for atom by atom fabrication, extraction of relevant physics and chemical behaviors on the single-atom levels, as well as mesoscopic studies of electromechanical and transport phenomena via scanning probe microscopy. Sergei has co-authored >600 publications, with a total citation of >30,000 and an h-index of >85. He is a fellow of MRS, APS, IoP, IEEE, Foresight Institute, and AVS; a recipient of the RMS medal for Scanning Probe Microscopy (2015); Blavatnik Award for Physical Sciences (2018), Presidential Early Career Award for Scientists and Engineers (PECASE) (2009); Burton medal of Microscopy Society of America (2010); 4 R&D100 Awards; and a number of other distinctions.

References

1. Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. (Academic Press; 2 edition, 2014).
2. Lambert, B. *A Student's Guide to Bayesian Statistics*. (SAGE Publications Ltd; 1 edition, 2018).
3. Martin, O. *Bayesian Analysis with Python: Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ, 2nd Edition*. (Packt Publishing, 2018).