Machine Learning for Scientific Data Analysis

Gabriele Scalia

Abstract Over the last few years, machine learning has revolutionized countless areas and fields. Nowadays, AI bears promise for analyzing, extracting knowledge, and driving discovery across many scientific domains such as chemistry, biology, and genomics. However, the specific challenges posed by scientific data demand to adapt machine learning techniques to new requirements. We investigate machine learning-driven scientific data analysis, focusing on a set of key requirements. These include the management of uncertainty for complex data and models, the estimation of system properties starting from low-volume and imprecise collected data, the support to scientific model development through large-scale analysis of experimental data, and the machine learning-driven integration of complementary experimental technologies.

1 Introduction

Managing scientific data poses unique and challenging problems. In particular, the volume, complexity and heterogeneity of the available data has led to an increasing relevance of data-driven methodologies to integrate, process and analyze it. To match the ambitious application goals, the complex and high-dimensional data sources and the quality issues characterizing this data, increasingly sophisticated methods have been proposed, with a central role played by artificial intelligence (AI). Machine learning (ML) and deep learning-based methods are revolutionizing applications which have been consolidated for years [3, 16, 17] or enabling totally new directions in tandem with new technological progresses [7].

This combination, even though extremely promising, is not without challenges, and requires an adaptation process on both sides [16]. On the one side, progresses in AI have led to rethinking consolidated fields such as drug discovery [3] and medicinal chemistry synthesis [17], just to name a few. On the other, the specificities

G. Scalia (✉)

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy
e-mail: gabriele.scalia@polimi.it

© The Author(s) 2022

L. Piroddi (ed.), *Special Topics in Information Technology*,
PoliMI SpringerBriefs, https://doi.org/10.1007/978-3-030-85918-3_10

115

of scientific data and individual fields demand the development of new methods or the adaptation of existing ones.

This work goes in the direction of exploring ML techniques for scientific data analysis, investigating how they can be tailored to the needs of the data in this context. In doing this, it focuses on a set of case studies spanning chemistry, biology and genomics. The common thread of this research is the set of challenges posed by scientific data, including *quality* limitations, experimental *uncertainty*, low *volume*, and *complementarity* of the available sources in terms of modalities, resolutions or measured ranges. Additionally, while complete and structured descriptions can exist in some cases, data ontologies are often only partial, evolving and semi-structured. In view of this, it appears relevant to investigate ML-driven methods and systems specifically tailored to scientific data management and analysis.

This is an interdisciplinary research that, focusing on the challenges emerging in fields such as chemistry, genomics and biomedical research, investigates ML-driven techniques to face a set of identified requirements: (1) the management of uncertainty for complex data and models such as deep neural networks (DNNs), (2) the estimation of system properties starting from imprecise, low-volume and evolving data, (3) the continuous validation of scientific models through large-scale comparisons with experimental data and (4) the unsupervised integration of multiple heterogeneous data sources related to different technologies to overcome individual technological limitations. Common to virtually all fields driven by experimental data, these requirements are faced through a set of case studies.

This chapter presents the main results included in the Ph.D. thesis [12]. In the following, each section focuses on one of the directions previously mentioned. Each direction is investigated in the context of a specific scientific application scenario, but results are general and can be easily extended to other domains.

2 Uncertainty Estimation and Domain of Applicability for Neural Networks

Experimental data are always characterized by some level of intrinsic variability and imprecision. Models trained on those data inherit that uncertainty and are also affected by another kind of uncertainty that comes from insufficient training samples, often difficult to quantify, especially for complex models such as DNNs. For these reasons, modeling uncertainty in DNNs has recently attracted great interest and currently represents a major research direction in the field [5]. Accounting for uncertainty does not only mean outputting a confidence score for a given input. It also means *changing the way predictions are made*, taking into account the concept of “unknown” during the training and/or inference phases, and *making sense of the resulting uncertainty estimates*, which need to reflect and satisfy some principles to be really useful and trustworthy for the users. The latter point is especially important since it has been shown that modern neural networks, even though really accurate,

are often “over-confident” in their output probabilities (e.g., a DNN could say that a certain image is 99% likely to be a cat, while the true observed confidence is much lower) [5]. Recent progresses on uncertainty estimation in DNN emerged in the computer vision field, mainly focusing on convolutional neural networks (CNNs). In that context, uncertainty prediction is often studied to overcome interpretability and safety limitations of modern computer vision applications such as autonomous driving [6].

When we consider scientific data and applications, uncertainty estimation assumes a unique relevance. Experimental datasets are often comparatively small (being costly to generate), sparse, and affected by various kinds of inherent imprecision such as experimental errors, lack of coherent ontologies, and misreporting [16]. On top of this, common requirements of scientific applications put particular emphasis on uncertainty. For example, *drug discovery* is strictly related to exploring the “uncharted” chemical space and, therefore, estimating the uncertainty over such predictions is crucial since there will always be a *knowledge boundary* beyond which predictions start to degrade. In such cases, uncertainty estimation becomes strictly related to the problem of defining a *domain of applicability* for a model [13].

We study this problem in the context of *molecular property prediction*, formally referred to as *Quantitative Structure-Activity Relationship* (QSAR). In the last few years, pioneering neural network architectures for QSAR, such as graph neural networks (GNNs), have been proposed. Such models, combined to an increasing availability of data and computational power, have led to state-of-the-art performance for this task. However, these models are still characterized by some key limitations, such as interpretability and generalization ability [16].

In this respect, we investigate how uncertainty can be modeled in DNNs, theoretically reviewing existing methods and experimentally testing them on GNNs for molecular property prediction. In parallel, we develop a framework to qualitatively and quantitatively evaluate the estimated uncertainties from multiple points of view. An overview of the methodology is shown in Fig. 1.

2.1 A Bayesian Graph Neural Network for Molecular Property Prediction

Uncertainty can be the result of inherent data noise or could be related to what the model does not yet know. These two kind of uncertainties—*aleatoric* and *epistemic*—can be combined to obtain the total predictive uncertainty of the model. We extend a GNN to model both uncertainty components.

When not explicitly modeled, the inherent observation noise is assumed constant for every observed molecule. However, this assumption does not hold in many realistic settings, where input-dependent noise needs to be modeled, such as chemistry applications. Data-dependent aleatoric uncertainty is referred to as *heteroscedastic* and its importance for DNNs has been recently highlighted [6]. Since aleatoric uncer-

Fig. 1 Overview of the methodology described in [13]. Reprinted with permission from G. Scalia, C. Grambow, B. Pernici, Y-P. Li, W. H. Green, *J. Chem. Inf. Model.* 2020, 60, 6, 2697–2717. Copyright 2020 American Chemical Society

tainty is a property of data, it can be learned directly from the data adapting the model and the loss function. However, aleatoric uncertainty does not account for epistemic uncertainty. This can be overcome by performing Bayesian inference, through the definition of a *Bayesian neural network*.

In a Bayesian neural network the weights of the model θ are *distributions* learned from training data $\mathcal{D}$, instead of point estimates, and therefore it is possible to predict the output distribution $\mathbf{y}$ of some new input $\mathbf{x}$ through the *predictive posterior distribution* $p(\mathbf{y} | \mathbf{x}, \mathcal{D}) = \int p(\mathbf{y} | \mathbf{x}, \theta) p(\theta | \mathcal{D}) d\theta$. Monte Carlo integration over M samples of the posterior distribution can approximate the intractable integral, however, obtaining samples from the true posterior is virtually impossible for DNNs. Therefore, an approximate posterior $q(\theta) \approx p(\theta | \mathcal{D})$ is introduced. A common technique to derive $q(\theta)$ is *variational inference* (VI). Approximate VI can scale up to the training-intensive large datasets/models of modern applications. Major examples of techniques of this type are MC-Dropout and ensembling-based methods. In [13], different techniques have been experimentally compared.

Experimental results on major public datasets [13] show that the computed uncertainty estimates allow correctly approximating the expected errors in many cases, and, in particular, when test molecules are comparatively similar with respect to training molecules. When this is not the case, uncertainty tends to be underestimated, but it still allows ranking test predictions by confidence. Moreover, experiments show how modeling both types of uncertainty is in general beneficial, and that the relative contribution of each uncertainty type to total uncertainty is dataset dependent. Additionally, it has been shown how modeling uncertainty has a consistent positive impact on model's accuracy.

The methodology, experimental results and additional analyses are detailed in [13] and in Chap. 3 of [12].

3 Machine Learning Estimation of System Properties from Uncertain, Low-Volume Data

ML methods are commonly used to learn a model of the data (in terms of inputs or inputs/outputs). Generally speaking, every ML model can be thought as a model of the observed data, and, therefore, as an estimator of a function of the data distributions [4]. This definition is general and includes both supervised and unsupervised tasks.

However, in many scientific settings the ultimate goal is not learning a model of the data, but using the data to learn something about the underlying system that has generated it. The way scientific data and ML techniques can be used in such contexts necessarily changes and is generally less investigated. This direction is characteristic of scientific domains, and it is also hindered by all the various types of quality limitations characterizing data in this context, such as imprecision and scarcity [9].

We explore the problem of estimating some properties of a system (e.g., a biological system) starting from imprecise, low-volume data. As a case study, we consider *biological systems* characterized by some kind of internal information exchanges, and we develop a ML-driven methodology to estimate the optimal way of transferring information given as input a set in-silico/in-vitro experiments for the system.

3.1 A Machine Learning-Driven Approach to Optimize Bounds on the Capacity of a Molecular Channel

We consider biological systems characterized by some kind of internal information exchange. Such exchange can be described from a communication point of view, allowing defining a *capacity* for the communication channel (which, in this case, is a molecular channel). We face the problem of estimating this capacity, which is strictly related to estimating the optimal way of transferring information in the system, only using a set of inputs/outputs for the system, that can be, for example, the result of in-silico/in-vitro experiments. This is particularly useful since, for many biological systems, an analytical or statistical description does not exist, but datasets of inputs/outputs can be easily obtained.

We propose a novel methodology that frames the estimation of the capacity as the optimization problem of finding an upper and a lower bound on the true value. The bounds are optimized starting from the data and using an evolutionary iterative algorithm. Being estimated from the data, the accuracy of the resulting interval is affected by the uncertainty and the volume of the available data. Therefore, particular emphasis has been placed on overcoming data scarcity and managing uncertainty in the available biological data. An overview of the methodology is shown in Fig. 2.

Two fundamental factors hinder the optimization procedure, and are addressed by the proposed methodology:

Fig. 2 Overview of the methodology presented in [10]. Using an iterative evolutionary optimization algorithm and a DNN-driven augmentation strategy, the proposed methodology converges to tight capacity bounds. © 2021 IEEE. Reprinted, with permission, from F. Ratti, G. Scalia, B. Pernici and M. Magarini, “A Data-driven Approach to Optimize Bounds on the Capacity of the Molecular Channel”, *GLOBECOM 2020 - 2020 IEEE Global Communications Conference, Taipei, Taiwan, 2020*, pp. 1–7

- The biological model is a *black-box*: an analytical characterization of its behavior is not known, and therefore an analytical formulation of the bounds does not exist. This prevents the usage of exact or gradient-based optimization algorithms. We overcome this issue with two complementary solutions: (1) evaluating the bounds functions directly on the data, and (2) using a derivative-free optimization algorithm (the Covariance Matrix Adaptation Evolution Strategy (CMA-ES)).
- Given the fact that biological simulations are heavily time consuming, we cannot rely on on-demand simulations to approximate any possible input distribution during the execution of iterative optimization algorithm. We address this issue by using a set of previously generated simulations as input of a DNN-based *data augmentation* module.

The proposed methodology is experimentally evaluated on a in-silico system composed of two prokaryotic cells. The methodology and the experimental results are detailed in [10] and in Chap. 4 of [12].

4 Data-Driven Validation and Development of Scientific Models

In many scientific settings, models are developed independently and externally with respect to the collected data and the empirical observations (for example, in principle-based models). These models are not designed directly from the data, but experimental data still have a key role in the development/validation/refinement cycle. Necessarily, the role of ML methods in this context changes. Indeed, in this context they primarily take a *validation* role, supporting the development cycle of the model leveraging the available data.

The sharing of scientific data has greatly increased in the last decade, leading to open repositories in many different scientific domains, also thanks to specific initiatives and guidelines [18]. However, it was recently discussed how data sharing practices have received far more study than has data reuse, and that *data sharing* and *open data* are not final goals in themselves, but the real benefit is in *data reuse*, which is “an understudied problem that requires much more attention if scientific investments are to be leveraged effectively” [8]. However, in order to reuse datasets from multiple sources, a series of challenges must be addressed.

We discuss the design of a framework to manage the development, validation and refinement cycle of scientific models taking advantage of large amounts of scientific data (experiments) extracted from the literature. As a case study we consider chemical kinetics models, which determine the reactivity of fuels and mixtures, but the approach taken is general and domain-agnostic.

4.1 Towards an Integrated Framework to Support Scientific Model Development

We investigate this direction from an information systems point of view, discussing the requirements and an architecture for the framework. Moreover, we develop a prototype to better analyze the framework’s requirements, that relies on a service-oriented architecture and provides a set of functions to import experiments and models, automatically run simulations and compute global validation indices/statistics.

The design of an integrated framework (see Fig. 3) to support model development through large-scale automatic validation on published experimental data requires addressing a set of use case [14, 15]. Those include: (1) *the acquisition of new experiments*, guaranteeing consistency, uniqueness and quality, (2) *the simulation of experiments through available models*, that requires automatically interpreting and handling data and models, (3) *cross comparisons and global validations*, including model/experiment validation through large-scale comparison, also supported by ML techniques, (4) *managing changes in models*, tracking and supporting the development through a continuous validation approach.

Fig. 3 Overview of the architecture of the SciExpeM framework, highlighting the main logical components. Reprinted from G. Scalia, M. Pelucchi, A. Stagni, A. Cuoci, T. Faravelli, B. Pernici, *Towards a scientific data framework to support scientific model development*, *Data Science*, 2 (1–2), IOS Press, 2020 with permission from IOS Press. The publication is available at IOS Press through <http://dx.doi.org/10.3233/DS-190017>

The SciExpeM framework [14] has been designed to support these use cases, automatizing the processes of acquiring, interpreting, simulating and cross-comparing models and experiments. SciExpeM provides *model validation* functionalities supported by large scale *data aggregation* and *data analysis* tools. Analysis tools include outlier detection, clustering and statistical correlations. Examples of outcomes that can be achieved through SciExpeM include: (1) finding potential errors in the experimental data through cross-comparisons with other experiments/models (also identifying possible causes for such errors), (2) tracking models, highlighting their performance over different versions with respect to a set of experiments and identifying critical areas, (3) identifying models/experiments remarkably different with respect to others available in the system to be validated.

The SciExpeM framework has been first studied in [15], and then largely extended in [14]. Chapter 5 of [12] details the main results.

5 Unsupervised Deep Learning-Driven Integration of Multiple Sources

Data integration has been a central research direction across many fields and from different perspectives for decades, with applications in countless areas. Nonetheless, integration in the context of scientific data is still considered a fundamental challenge to be addressed. For example, the “integration of single-cell data across samples, experiments, and types of measurement” has been very recently highlighted as one of the main challenges in single-cell data science [7].

One limitation of traditional integration techniques is that of being mostly rule-driven. Instead, data-driven (including ML-driven) methodologies can address the lack of ontologies and complete descriptions of the underlying phenomena, not requiring strong a priori assumptions about the available data sources (in terms of quality, available modalities, etc.). Moreover, they can help exploit *complementary strengths* of the available data sources and enable context-aware analysis.

In the scientific domain, data integration is often an *enabler* for other activities. Integration can help overcoming data scarcity and quality issues, thus indirectly improving downstream data-driven applications which rely on large, high-quality datasets [16]. Integration can also allow overcoming technological limitations hindering next-generation applications. For example, creating comprehensive and multi-scale biological atlases at single cell resolution of the human body is currently recognized as next frontier to understand cellular basis of health and disease [11]. However, a single experimental technology for doing that currently does not exist: multiple technologies capturing different aspects, scales and modalities are available.

We investigate this problem designing a novel methodology, named *Tangram* [1], to (1) integrate transcriptomes of cells, overcoming limitations of existing single-cell RNA sequencing technologies, and (2) relate cellular features to the histological and anatomical scales through integration in the *anatomical space*. We show how, starting from complementary experimental datasets obtained through different technologies characterized by different limitations, it is possible to address the individual technological limitations via integration. The proposed methodology is unsupervised and requires minimal domain knowledge.

5.1 Machine Learning-Driven Alignment of Spatially-Resolved Whole Transcriptomes with Tangram

We tackle integration challenges posed by the creation of high-resolution cell atlases from two complementary perspectives: (1) learning an alignment between experimental data measured through different technologies and (2) learning an anatomical manifold from a pre-existing atlas that allows the integration of the new data. For example, given a tissue for which we have collected single-cell RNA-sequencing data, Spatial Transcriptomics data (not necessarily at single-cell resolution) and his-

tological images, we can (1) integrate single cells and Spatial Transcriptomics data, obtaining a high-quality cell-resolution map of all the genes for the tissue and (2) integrate the obtained transcriptomes to an existing organ-level or human-level cell atlas using the histological images. For both these directions the methodology proposed is fully data-driven, unsupervised, does not require previously defined rules and does not make strong assumptions about the available data sources. For example, the method is flexible with respect to the number and the type of measured genes, the available experimental sources, and the quality of the data (indeed, Tangram can improve data quality through integration).

Intuitively, the way Tangram achieves integration resembles a puzzle game. Tangram uses single cell RNA-sequencing data as “puzzle pieces” to align in space to match “the shape” of the spatial data. From the learned mapping function, Tangram can (1) expand from a measured subset of genes to genome wide profiles, (2) correct low-quality spatial measurements, (3) map and show the location of cells of different types in space, (4) infer the mixture of cells collectively measured through low resolution measurements, and (5) align multi-modal data at single cell resolution using transcriptomics data as a bridge. Technically, Tangram is based on the optimization through gradient descent of an “alignment” function that allows obtaining a “mapping”, through which two complementary transcriptomics datasets are aligned. The integration at the histological and anatomical scales is achieved building a latent space as a proxy of a similarity metric through a Siamese neural network with a semantic segmentation algorithm.

Through a large set of experiments using various datasets measured with different technologies, we show how Tangram can be used to improve the resolution, the throughput, the quality and the available modalities of the starting datasets via integration. As a final case study we show how we can learn a histological and anatomical integrated atlas of the somatomotor area of the healthy adult mouse brain starting from publicly available datasets and atlases [1]. Additionally, Tangram has been used in conjunction with other computational methods to provide a global picture of the regulatory mechanisms that govern cellular diversification in the mammalian neocortex [2]. Details about the Tangram method and the experimental results can be found in [1] and in Chap. 6 of [12].

6 Conclusion

In this work we have investigated different facets of knowledge extraction from scientific data, focusing on ML methods and the specific challenges posed by scientific and experimental data analysis. Over the last few years, machine learning has been a game-changing technology across countless areas and fields. However, the specific features of scientific data demand to adapt these techniques to new requirements and roles. This research investigated several of these requirements, including the role of experimental uncertainty in data modeling, the ML-driven inference of biological systems properties, the role of ML in supporting data-driven scientific

model development, and the ML-driven integration of complementary experimental technologies. The choice of considering multiple application scenarios (spanning chemistry, biology, and genomics) in this work derives from the objective of exploring these challenges under multiple point of views. Research on this topic is growing at a staggering rate and is opening up directions for many areas in ML.

References

1. T. Biancalani, G. Scalia, et al., *Deep Learning and Alignment of Spatially-Resolved Single Cell Transcriptomes with Tangram*. Accepted for publication in Nature Methods (2021)
2. D.J. Di Bella, E. Habibi, R.R. Stickels, G. Scalia, J. Brown, P. Yadollahpour, S.M. Yang, C. Abbate, T. Biancalani, E.Z. Macosko, F. Chen, A. Regev, P. Arlotta, Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* (2021)
3. E. Gawehn, J.A. Hiss, G. Schneider, Deep learning in drug discovery. *Mol. Informatics* **35**(1), 3–14 (2016)
4. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT press, 2016)
5. C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in *Proceedings of the 34th International Conference on Machine Learning, ICML'17*, (2017), pp. 1321–1330
6. A. Kendall, Y. Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision?, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, (2017), pp. 5580–5590
7. D. Lähnemann, J. Köster, E. Szczurek, D.J. McCarthy, S.C. Hicks, M.D. Robinson, C.A. Vallejos, K.R. Campbell, N. Beerenwinkel, A. Mahfouz et al., Eleven grand challenges in single-cell data science. *Genome Biol.* **21**(1), 1–35 (2020)
8. I.V. Pasquetto, B.M. Randles, C.L. Borgman, On the reuse of scientific data. *Data Sci. J.* **16**(8), 1–9 (2017)
9. B. Pernici, F. Ratti, G. Scalia, *About the Quality of Data and Services in Natural Sciences* (Springer International Publishing, Cham, 2021), pp. 236–248
10. F. Ratti, G. Scalia, B. Pernici, M. Magarini, A data-driven approach to optimize bounds on the capacity of the molecular channel, in *2020 IEEE Global Communications Conference (GLOBECOM)* (IEEE, 2020), pp. 1–7
11. A. Regev, S.A. Teichmann, E.S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy et al., Science forum: the human cell atlas. *Elife* **6**, e27041 (2017)
12. G. Scalia, Machine Learning-Driven Integration, Knowledge Extraction and Uncertainty Management for Scientific Data. Ph.D. thesis, Politecnico di Milano (2020)
13. G. Scalia, C.A. Grambow, B. Pernici, Y.P. Li, W.H. Green, Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Information Modeling* **60**(6), 2697–2717 (2020)
14. G. Scalia, M. Pelucchi, A. Stagni, A. Cuoci, T. Faravelli, B. Pernici, Towards a scientific data framework to support scientific model development. *Data Sci.* **2**(1–2), 245–273 (2019)
15. G. Scalia, M. Pelucchi, A. Stagni, T. Faravelli, B. Pernici, Storing combustion data experiments: new requirements emerging from a first prototype, in *Semantics, Analytics, Visualization, 3rd International Workshop, SAVE-SD 2017*, Revised Selected Papers, LNCS, vol. 10959 (Springer International Publishing, Cham, 2018), pp. 138–149
16. P. Schneider, W.P. Walters, A.T. Plowright, N. Sieroka, J. Listgarten, R.A. Goodnow, J. Fisher, J.M. Jansen, J.S. Duca, T.S. Rush, et al., Rethinking drug design in the artificial intelligence era, in *Nature Reviews Drug Discovery* (2019), pp. 1–12

17. T.J. Struble, et al., Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J. Med. Chem.* (2020)
18. M.D. Wilkinson et al., The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**(160018), 1–9 (2016)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.