Sionna: An Open-Source Library for Next-Generation Physical Layer Research

Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, Alexander Keller

Abstract—Sionna™ is a GPU-accelerated open-source library for link-level simulations based on TensorFlow. It enables the rapid prototyping of complex communication system architectures and provides native support for the integration of neural networks. Sionna implements a wide breadth of carefully tested state-of-the-art algorithms that can be used for benchmarking and end-to-end performance evaluation. This allows researchers to focus on their research, making it more impactful and reproducible, while saving time implementing components outside their area of expertise. This white paper provides a brief introduction to Sionna, explains its design principles and features, as well as future extensions, such as integrated ray tracing and custom CUDA kernels. We believe that Sionna is a valuable tool for research on next-generation communication systems, such as 6G, and we welcome contributions from our community.

I. INTRODUCTION

WHILE 5G is deployed around the globe and the work areas for its future evolution called 5G Advanced (Release 18) have just been approved [1], researchers in academia and industry have already started to define visions and key technologies for 6G, e.g., [2]. One recurring vision in many papers is that of creating digital twin worlds to connect physical and biological entities and to enable new mixed-reality experiences such as the Metaverse [3].

Some of the most prominent 6G research topics are communication in the Terahertz band [4], cell-free, holographic, and very large aperture multiple-input multiple-output (MIMO) [5], unmanned aerial vehicle (UAV) and satellite communication [6], machine learning (ML)-based air interface design [7], semantic communication [8], reconfigurable intelligent surfaces (RIS) [9], sensing and localization [10], digital twins [11], computer vision for wireless [12], as well as federated learning over wireless [13]. Interestingly, research on most of the above topics requires one or more of the following:

- **Data from specific radio environments**: While the widely used stochastic channel models, e.g., [14], are well suited to capture the behavior of a certain class of environment, they cannot be used for any application that requires the simulation of a specific environment, e.g., the optimal configuration of a particular RIS. Radar-based sensing and localization as well as computer vision-aided applications only work if there is a spatially consistent correspondence between physical location and channel impulse response (or visual input). This can only be achieved through either ray tracing or extensive measurement campaigns.
- **Native integration of ML**: ML and, particularly, neural networks (NNs) are expected to play an increasingly important role for the implementation of transceiver algorithms and maybe even the 6G air interface design [1], [7], [15]. Research on such topics does not only benefit tremendously from a tight integration of ML, e.g., [16], [17], and link-level simulation tools, but also from automatic gradient computation through the entire system which allows for seamless integration of NNs.
- **Very high detail or scale**: Link-level simulations of 6G systems require unprecedented modeling accuracy and scale. While cell-free or very large aperture MIMO systems need the computation of a very large number of signal propagation paths, Terahertz communication systems have complex channel models accounting for nonlinear hardware impairments. Moreover, the full potential of ML-enhanced algorithms will only be unleashed with very realistic simulations that account for detail that has been ignored in the past due to prohibitive simulation complexity and for mathematical tractability.

Due to the reasons above, we firmly believe that the breakthroughs required for 6G can only be achieved if our community has access to new tools for physical layer research. That is why we have started to develop a new and open-source link-level simulator called Sionna.¹

II. MOTIVATION & BACKGROUND

A. Why we decided to create Sionna

Sionna² is written by researchers for researchers in communications and aims at making our work more efficient:

- **Rapid prototyping**: Sionna provides a high-level Python application programming interface (API) to rapidly model complex communication systems from end-to-end while allowing one to easily adapt the parts a new research idea is about. Graphic processing unit (GPU) acceleration makes it super fast and enables interactive exploration, e.g., in Jupyter notebooks [18].
- **Benchmarking against the state of the art**: Sionna provides many carefully tested standard processing blocks and state-of-the-art algorithms that can be used for performance benchmarking. This reduces the time spent implementing auxiliary components that one’s research is not about.
- **Realistic industry-grade evaluations**: Experts in one domain, say channel estimation, do not necessarily have

¹https://github.com/NVlabs/sionna
²Sionna is a goddess in Irish mythology and the namesake of the River Shannon, see https://en.wikipedia.org/wiki/Shannon_(given_name).
A. Design principles

One of the key differences between Sionna and alternative link-level simulators is that all algorithms are implemented using high-dimensional Tensors, e.g., of shape [batch_size, num_tx,..., fft_size] where the first dimension is the batch size, representing different independent Monte Carlo trials that are executed in parallel. For-loops are generally avoided. This representation is borrowed from the field of deep learning and leads to “embarrassingly parallel” workloads that can be efficiently executed on GPUs. If no GPU is available, Sionna will also run on (multiple) central processing units (CPUs).

Another design principle of Sionna is that all components are implemented as independent Keras layers [17]. This has the advantages that (i) complex system models can be constructed by simply connecting the desired layers, (ii) components can be easily replaced by NNs, and (iii) gradients are automatically backpropagated through an entire system, which is the key enabler for end-to-end learning of new artificial intelligence (AI)-defined air interfaces. Sionna also eliminates the need for different tools for data generation, training, and evaluation.

• **Native support of NNs:** Sionna enables seamless integration of NNs in the physical layer signal processing chain. As most building blocks are differentiable, gradients can be backpropagated through an entire system, which is the key enabler for end-to-end learning of new artificial intelligence (AI)-defined air interfaces. Sionna also eliminates the need for different tools for data generation, training, and evaluation.

• **Addressing 6G research needs:** Sionna will allow for the use of ray tracing, datasets, and generative models instead of stochastic channel models to enable research on many novel topics, such as joint communication and sensing, vision-aided wireless communications, intelligent reflecting surfaces, semantic communications, and digital twins. Also, a THz channel model is under development. Community contributions to components are welcome.

• **Reproducibility:** With Sionna, it is straight-forward to make research reproducible by others. We encourage all users to publish their Sionna-based code along with their research papers and contribute novel components so that they can be reused by others. This will also make it easier to compare algorithms under the same conditions.

**B. Related open-source software (OSS)**

Commercial software products, such as Matlab [19], play an important role for simulation-based research in our field. There is also a growing body of well-maintained OSS projects serving different purposes, such as channel simulation [20], [21], link-level simulations [22]–[24], system-level simulations [25], network simulations [26], [27], and hardware experimentation [28], [29]. There are also several open experimental platforms, such as the Platforms for Advanced Wireless Research (PAWR) in the US [30] and the OneLab Platforms [31] in Europe. These allow for reproducible large-scale experiments in controlled and/or realistic environments. Not all the mentioned OSS, e.g., [20], [21], [24], accept contributions from external parties though.

Sionna is a tool for link-level simulations with in-built channel simulator. It has hence some similarities with Quadriga [20] and NYUSIM [21] for the generation of channel impulse responses, as well as with the Vienna 5G Link-Level Simulator [22] and HermesPy [24] for physical layer processing. It differs, however, through its end-to-end differentiability, native support of NNs and GPUs, as well as ray tracing capabilities. Sionna can be scaled to large multi-GPU setups to simulate, e.g., realistic multi-cell Massive MIMO systems.

III. A SIONNA PRIMER

Sionna is written in Python using TensorFlow [16] and the Keras API [17]. Exceptions are the custom CUDA kernels and ray tracing capabilities that will be described in Sec. IV.

**Channel models:**

- Additive white Gaussian noise (AWGN) channel
- Flat-fading channel models with antenna correlation
- 3GPP 38.901 TDL, CDL, UMa, UMi, RMa models [14]
- Import of channel impulse response from datasets
- Channel output computed in time or frequency domain

**Forward error correction (FEC):**

- 5G LDPC codes including rate matching [33]
- 5G Polar codes including rate matching [33]
- Cyclic redundancy check (CRC) [33]
- Reed-Muller & Convolutional codes
- Interleaving & Scrambling
- Belief propagation (BP) decoder and variants
- SC, SCL, and SCL-CRC Polar decoders
- Viterbi decoder
- Demapper with prior
- EXIT chart simulations
- Import of parity-check matrices in alist format

**III. A SIONNA PRIMER**

Sionna is written in Python using TensorFlow [16] and the Keras API [17]. Exceptions are the custom CUDA kernels and ray tracing capabilities that will be described in Sec. IV.
MIMO processing:
- Multiuser & multicell MIMO support
- 3GPP 38.901 & custom antenna arrays/patterns
- Zero forcing (ZF) precoding
- Minimum mean squared error (MMSE) equalization

Orthogonal frequency-division multiplexing (OFDM):
- OFDM modulation & demodulation
- Cyclic prefix insertion & removal
- Flexible 5G slot-like frame structure
- Arbitrary pilot patterns
- LS channel estimation & Nearest neighbor interpolation

Despite the fact that Sionna supports 5G-compliant channel codes and channel models, it does not try by any means to be a 5G-compliant link-level simulator. It rather allows researchers to test their algorithms under realistic conditions and to compare to state-of-the-art algorithms that are widely accepted in the industry, while having a maximum degree of flexibility and freedom. The current set of features should be seen as a good starting point that one can build on to rapidly prototype and evaluate new ideas.

C. Hello, World!

The Sionna webpage provides an extensive set of tutorial notebooks (https://nvlabs.github.io/sionna/tutorials.html) that cover in detail the various ways the library can be used. In order to get a first impression of how developing with Sionna feels like, we provide and discuss a few short code examples.

Listing 1 shows a Sionna “Hello, World!” example in which the transmission of \( \text{batch\_size} \) LDPC codewords over an AWGN channel using 16QAM modulation is simulated. This example shows how Sionna layers are instantiated and then immediately applied to a previously defined tensor. For example, in line 8, the LDPC5GEncoder is instantiated with the parameters \( n \) and \( k \) and then applied to the tensor \( b \) which contains randomly generated information bits. This coding style follows the functional API of Keras.

In Listing 2, we have made the Constellation trainable and replaced the Demapper by a NeuralDemapper which is defined in Listing 3. What happens under the hood is that the tensor defining the constellation points has become a trainable TensorFlow Variable. It can now be tracked together with the weights of the NN of the NeuralDemapper by TensorFlow’s automatic differentiation feature, see [34]. One could then define an adequate loss function (line 7), such as the total binary cross-entropy, and compute the gradient of this loss with respect to all trainable variables. This gradient may then be used to optimize jointly the constellation points and the NeuralDemapper through stochastic gradient descent. This concept is sometimes referred to as end-to-end learning [35] and can be applied to very complex system models.

Listing 3 shows the definition of a NeuralDemapper as a Keras layer which was used in Listing 2 to replace a traditional Demapper. This simple combination of NN components and traditional algorithms together with automatic gradient computation is one of the key advantages of Sionna for ML-based research.

D. Limitations

As Sionna parallelizes simulations over batches, the available GPU memory can quickly become the bottleneck. This can be circumvented by either reducing the batch size or distributing the model or batch across multiple GPUs.

Algorithms with complex conditional logic, e.g., where examples in a batch can be treated differently, do not lend themselves to be written efficiently using TensorFlow’s Python API. In this case, one can resort to implementing custom TensorFlow operations in C++ and CUDA (see Sec. IV-A). At the cost of execution speed, it is also possible to write such algorithms in native Python code and wrap them as a TensorFlow operations.
Listing 4. Sionna ray tracing example.

def load_scene(sionna.rt.scene.munich):
    # Load integrated scene
    scene = load_scene(sionna.rt.scene.munich)
    # Compute propagation paths
    paths = scene.compute_paths()
    # Compute coverage map
    cm = scene.coverage_map()
    # Render scene with coverage map and paths
    scene.render(paths=paths, coverage_map=cm)
    # Convert paths to channel impulse responses
    a, tau = p2c(paths.as_tuple())
    p2c = Paths2CIR(sampling_frequency=1e6, scene=scene)

E. Contributing

Sionna is released under the Apache 2.0 license. We have chosen this license so that users do not need to worry about possible patent infringements and litigation. We accept and welcome contributions from external parties through pull request on the Sionna GitHub repository. The only requirement is that all contributors “sign-off” their commits with the Developer Certificate of Origin (DCO) [36].

IV. Future extensions

A. Custom CUDA kernels

We have chosen Python and TensorFlow for writing most of Sionna due to their popularity and simplicity. However, not all signal processing algorithms and simulation routines can be easily expressed using tensors and compositions of existing TensorFlow operations. This is generally the case, whenever an algorithm might execute differently for each example in a batch, e.g., due to if-else statements, or requires complex indexing. Examples comprise the Polar SCL and min-sum BP decoder, components of the 3GPP channel models, as well as the convolution with a time-varying channel impulse response. Moreover, some algorithms expressed as the composition of TensorFlow operations might not lead to acceptable performance or memory requirements.

As an alternative, one can implement such algorithms without the constraints of the TensorFlow API as custom TensorFlow operations [37] using C++ and XLA [32] or CUDA [38] to benefit from GPU acceleration. We found the resulting code to be significantly more convenient, better readable, and less prone to errors. The source code of a custom operation is compiled into a binary shared object that can be loaded by TensorFlow from the Python code. The custom operation can then be called from the Python TensorFlow API in the same way as any other operation. Although tensors are used as inputs and outputs to custom operations, any type of data structure can be used in the internal C++ implementation with CUDA. Moreover, the CUDA-X libraries [39] can be leveraged to further benefit from efficient GPU-accelerated algorithms. If required, gradients can be specified to make a custom operation differentiable.

B. Ray tracing

As mentioned earlier, many 6G topics, such as RIS or integrated sensing and communications, require the simulation of a specific radio environment in a physically-based manner which cannot be done with stochastic channel models. For some applications, also the visual representation of a scene is required, see, e.g., [40]. To address these needs, Sionna RT will soon bring both of these capabilities to Sionna so that one can render a scene, get the desired channel impulse responses, and use them immediately for link-level simulations or other applications.

Listing 4 demonstrates the simplicity of invoking high performance ray tracing from within a Jupyter notebook. First, an integrated scene is loaded. Then, propagation paths between all transmitters and receivers are computed. Their configuration is not shown in this listing. Next, a coverage map is generated before it is rendered in the scene together with the propagation paths (see Fig. 1). Finally, channel impulse responses between all defined transmitters and receivers are computed. These can be directly used for link-level simulations instead of a stochastic channel model. Beyond that simplicity, one can manipulate many scene parameters, such as material properties, from within Python.

V. Conclusions

The next decade will be marked by groundbreaking research on disruptive technologies for 6G whose exploration poses various challenges for computer simulations. For this reason, we have started the development of Sionna, a differentiable open-source link-level simulator with native integration of NNs and full GPU acceleration. Soon, Sionna will have an integrated ray tracer for scene rendering and wave propagation. We are excited about the possibilities Sionna offers and hope for wide adoption and numerous contributions from our community.

REFERENCES

[1] X. Lin, “An Overview of 5G Advanced Evolution in 3GPP Release 18,” arXiv preprint arXiv:2201.01358, Jan. 2022.
[2] M. A. Uusitalo, M. Ericson, B. Richerzhagen, E. U. Soykan, P. Ruggeland, G. Fettweis, D. Sabella, G. Wikström, M. Boldi, M.-H. Hamon et al., “Hexa-X the European 6G flagship project,” in Proc. Joint Europ. Conf. Netw. Commun. & 6G Summit (EuCNC/6G Summit), Porto, Portugal, Jun. 2021, pp. 580–585.
Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, “Federated Learning in 6G Communication Networks: A Novel Multi-Cell Channel Model with Time Evolution for Enabling Virtual Field Trials,” IEEE Trans. Antennas Propag., vol. 62, no. 6, pp. 3242–3256, 2014, accessed Jan. 2022. [Online]. Available: https://wireless.engineering.nyu.edu/nyusim/
communications/.

G. Gezici, A. Garcia-Rodriguez, M. A. Azari, A. Loranzo, M. Mezzavilla, S. Chatzinotas, Y. Chen, S. Rangan, and M. Di Renzo, “What Will Be the Future of UAV Cellular Communications Be? A Flight from 5G to 6G,” arXiv preprint arXiv:2105.04842, May 2021.

J. Hoydis, F. A. Aoudia, A. Valsecchi, and H. Viswanathan, “Toward a 6G AI-Native Air Interface,” IEEE Commun. Mag., vol. 59, no. 5, pp. 76–81, May 2021.

M. Kountouris and N. Pappas, “Semantics-Empowered Communication for Networked Intelligent Systems,” IEEE Commun. Mag., vol. 59, no. 6, pp. 96–102, Jun. 2021.

M. D. Renzo, M. Debbah, D.-T. Phan-Huy, A. Zappone, M.-S. Alouini, C. Yuen, V. Sciancalepore, J. Hoydis, H. Gacanin, J. D. Rosny, A. Bournec, G. Lerosey, and M. Fink, “Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come,” EURASIP J. Wireless Commun. Netw., vol. 2019, no. 1, p. 129, May 2019.

A. Bourdoux, A. N. Barreto, B. van Liempd, C. de Lima, D. Dardari, T. Dittrich, T. Suutala, J. Saloranta, M. Guillaud, M. Isomursu, M. Valkama, M. K. Aziz, R. Berkvens, T. Sanguanpuak, T. Svensson, and Y. Miao, “6G White Paper on Localization and Sensing,” University of Oulu: 6G Research Visions, Jun. 2020.

P. Almasan, M. Ferrio-Lalmèns, J. Paillaiste, J. Suutala, J. Saloranta, M. Guillaud, M. Isomursu, M. Valkama, M. K. Aziz, R. Berkvens, T. Sanguanpuak, T. Svensson, and Y. Miao, “6G White Paper on Localization and Sensing,” University of Oulu: 6G Research Visions, Jun. 2020.

T. Nishio, Y. Koda, J. Park, M. Bennis, and K. Doppler, “When wireless communications meet computer vision in beyond 5G,” IEEE Commun. Standards Mag., vol. 5, no. 2, pp. 76–83, Jun. 2021.

Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, “Federated learning for 6G communications: Challenges, methods, and future directions,” China Communications, vol. 17, no. 9, pp. 105–118, Sep. 2020.

ETS1, “ETSI TR 138 901 V16.1.0: Study on channel model for frequencies from 0.5 to 100 GHz,” ETSI, Tech. Rep., Nov. 2020. [Online]. Available: https://www.etsi.org/deliver/etsi_frn/138901/138901v16.01.00_60/fr_138901v160100p0.pdf

S. Alli, W. Saal, and M. Steinbach, Eds., White Paper on Machine Learning in Wireless Communications Networks. University of Oulu, 2020, no. 7. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2015060272673

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, A. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, accessed Jan. 2022. [Online]. Available: https://www.tensorflow.org/

F. Chollet et al., “Keras,” https://keras.io, 2015, accessed Jan. 2022.

T. Kluiver, B. Ragan-Kelley, F. Perez, B. Granger, M. Bussommer, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, A. Avila, S. Abdalla, and C. Willing, “Jupyter Notebooks – a publishing format for reproducible computational workflows,” in Positioning and Power in Academic Publishing: Players, Agents and Agendas, F. Loizides and B. Schmidt, Eds. IOS Press, 2016, pp. 87–90. [Online]. Available: https://jupyter.org

T. M. Inc., “Matlab,” https://www.mathworks.com/products/matlab.html, accessed Jan. 2022.

S. Sun, G. R. MacCartney, and T. S. Rappaport, “A novel millimeter-wave channel simulator and applications for 5G wireless communications,” in Proc. IEEE Vehicular Technology Conf. (VTC2019-Spring), May 2020, pp. 1–5.