

PyGeNN: A Python library for GPU-enhanced neural networks

James C Knight^{1,*}, Anton Komissarov^{2,3}, Thomas Nowotny¹

¹*Centre for Computational Neuroscience and Robotics, School of Engineering and Informatics, University of Sussex, Brighton, United Kingdom*

²*Bernstein Center for Computational Neuroscience Berlin, Germany*

³*Technische Universität Berlin, Germany*

Correspondence*:

James C Knight

J.C.Knight@sussex.ac.uk

2 ABSTRACT

3 More than half of the Top 10 supercomputing sites worldwide use GPU accelerators and they
4 are becoming ubiquitous in workstations and edge computing devices. GeNN is a C++ library for
5 generating efficient spiking neural network simulation code for GPUs. However, until now, the full
6 flexibility of GeNN could only be harnessed by writing model descriptions and simulation code in
7 C++. Here we present PyGeNN, a Python package which exposes all of GeNN's functionality to
8 Python with minimal overhead. This provides an alternative, arguably more user-friendly, way
9 of using GeNN and allows modellers to use GeNN within the growing Python-based machine
10 learning and computational neuroscience ecosystems. In addition, we demonstrate that, in both
11 Python and C++ GeNN simulations, the overheads of recording spiking data can strongly affect
12 runtimes and show how a new spike recording system can reduce these overheads by up to
13 10×. Using the new recording system, we demonstrate that by using PyGeNN on a modern GPU,
14 we can simulate a full-scale model of a cortical column faster even than real-time neuromorphic
15 systems. Finally, we show that long simulations of a smaller model with complex stimuli and
16 a custom three-factor learning rule defined in PyGeNN can be simulated almost two orders of
17 magnitude faster than real-time.

18 **Keywords:** GPU, high-performance computing, parallel computing, benchmarking, computational neuroscience, spiking neural
19 networks, Python

1 INTRODUCTION

20 A wide range of spiking neural network (SNN) simulators are available, each with their own application
21 domains. NEST (Gewaltig and Diesmann, 2007) is widely used for large-scale point neuron simulations
22 on distributed computing systems; NEURON (Carnevale and Hines, 2006) and Arbor (Akar et al., 2019)
23 specialise in the simulation of complex multi-compartmental models; NeuroKernel (Givon and Lazar, 2016)
24 is focused on emulating fly brain circuits using Graphics Processing Units (GPUs); and CARLsim (Chou
25 et al., 2018), ANNarchy (Vitay et al., 2015), Spice (Bautembach et al., 2021), NeuronGPU (Golosio et al.,
26 2021) and GeNN (Yavuz et al., 2016) use GPUs to accelerate point neuron models. For performance
27 reasons, many of these simulators are written in C++ and, especially amongst the older simulators,
28 users describe their models either using a Domain-Specific Language (DSL) or directly in C++. For

programming language purists, fully custom DSLs such as the HOC network description language in NEURON (Carnevale and Hines, 2006) or the NestML (Plotnikov et al., 2016) neuron modelling language may be elegant solutions and, for simulator developers, using C++ directly and not having to add bindings to another language is convenient. However, both choices act as a barrier to potential users. Therefore, with both the computational neuroscience and machine learning communities gradually coalescing towards a Python-based ecosystem with a wealth of mature libraries for scientific computing (Hunter, 2007; Van Der Walt et al., 2011; Millman and Aivazis, 2011), exposing spiking neural network simulators to Python with minimal domain specific modifications seems like a pragmatic choice. NEST (Eppler et al., 2009), NEURON (Hines et al., 2009) and CARLsim (Balaji et al., 2020) have all taken this route and now all offer Python interfaces. Furthermore, newer simulators such as Arbor and Brian2 (Stimberg et al., 2019) have been designed from the ground up with a Python interface.

Our GeNN simulator can already be used as a backend for the Python-based Brian2 simulator (Stimberg et al., 2019) using the Brian2GeNN interface (Stimberg et al., 2020) which modifies the C++ backend “cpp_standalone” of Brian 2 to generate C++ input files for GeNN. As for cpp_standalone, initialisation of simulations is mostly done in C++ on the CPU and recording data is saved into binary files and re-imported into Python using Brian 2’s native methods. While we have recently demonstrated some very competitive performance results (Knight and Nowotny, 2018, 2020) using GeNN in C++, and through the Brian2GeNN interface (Stimberg et al., 2020), GeNN could so far not be used directly from Python and it is not possible to expose all of GeNN’s unique features through the Brian2 API. Specifically, GeNN not only allows users to easily define their own neuron and synapse models but, also ‘snippets’ for offloading the potentially costly initialisation of model parameters and connectivity onto the GPU. Additionally, GeNN provides a lot of freedom for users to integrate their own code into the simulation loop. In this paper we describe the implementation of PyGeNN – a Python package which aims to expose the full range of GeNN functionality with minimal performance overheads. Unlike in the majority of other SNN simulators PyGeNN allows defining bespoke neuron and synapse models directly from Python without requiring users to extend the underlying C++ code. Below, we demonstrate the flexibility and performance of PyGeNN in two scenarios where minimising performance overheads is particularly critical.

- In a simulation of a large, highly-connected model of a cortical microcircuit (Potjans and Diesmann, 2014) with small simulation timesteps. Here the cost of copying spike data off the GPU from a large number of neurons every timestep can become a bottleneck.
- In a simulation of a much smaller model of Pavlovian conditioning (Izhikevich, 2007) where learning occurs over 1 h of biological time and stimuli are delivered – following a complex scheme – throughout the simulation. Here any overheads are multiplied by a large number of timesteps and copying stimuli to the GPU can become a bottleneck.

Using the facilities provided by PyGeNN, we show that both scenarios can be simulated from Python with only minimal overheads over a pure C++ implementation.

2 MATERIALS AND METHODS

2.1 GeNN

GeNN (Yavuz et al., 2016) is a library for generating CUDA (NVIDIA et al., 2020) code for the simulation of spiking neural network models. GeNN handles much of the complexity of using CUDA directly and automatically performs device-specific optimizations so as to maximize performance. GeNN consists of a

69 main library – implementing the API used to define models as well as the generic parts of the code generator
 70 – and an additional library for each backend (currently there is a reference C++ backend for generating
 71 CPU code and a CUDA backend. An OpenCL backend is under development). Users describe their model
 72 by implementing a `modelDefinition` function within a C++ file. For example, a model consisting of 4
 73 Izhikevich neurons with heterogeneous parameters, driven by a constant input current might be defined as
 74 follows:

```
75 void modelDefinition(ModelSpec &model)
76 {
77     model.setDT(0.1);
78     model.setName("izhikevich");
79
80     NeuronModels::IzhikevichVariable::VarValues popInit(
81         -65.0, -20.0, uninitialisedVar(), uninitialisedVar(),
82         uninitialisedVar(), uninitialisedVar());
83
84     model.addNeuronPopulation<NeuronModels::IzhikevichVariable>(
85         "Pop", 4, {}, popInit);
86
87     model.addCurrentSource<CurrentSourceModels::DC>(
88         "CS", "Pop", {10.0}, {});
89 }
```

90 The `genn-buildmodel` command line tool is then used to compile this file; link it against the main GeNN
 91 library and the desired backend library; and finally run the resultant executable to generate the source code
 92 required to build a simulation dynamic library (a .dll file on Windows or a .so file on Linux and Mac). This
 93 dynamic library can then either be linked against a simulation loop provided by the user or dynamically
 94 loaded by the user's simulation code. To demonstrate this latter approach, the following example uses the
 95 `SharedLibraryModel` helper class supplied with GeNN to dynamically load the previously defined model,
 96 initialise the heterogenous neuron parameters and print each neuron's membrane voltage every timestep:

```
97 #include "sharedLibraryModel.h"
98
99 int main()
100 {
101     SharedLibraryModel<float> model("./", "izhikevich");
102     model.allocateMem();
103     model.initialize();
104     float *aPop = model.getScalar<float>("aPop");
105     float *bPop = model.getScalar<float>("bPop");
106     float *cPop = model.getScalar<float>("cPop");
107     float *dPop = model.getScalar<float>("dPop");
108     aPop[0] = 0.02; bPop[0] = 0.2; cPop[0] = -65.0; dPop[0] = 8.0;
109     aPop[1] = 0.1; bPop[1] = 0.2; cPop[1] = -65.0; dPop[1] = 2.0;
110     aPop[2] = 0.02; bPop[2] = 0.2; cPop[2] = -50.0; dPop[2] = 2.0;
111     aPop[3] = 0.02; bPop[3] = 0.2; cPop[3] = -55.0; dPop[3] = 4.0;
112     model.initializeSparse();
113
114     float *vPop = model.getScalar<float>("VPop");
115     while(model.getTime() < 200.0f) {
```

```

116     model.stepTime();
117     model.pullVarFromDevice("Pop", "V");
118     printf("%f, %f, %f, %f, %f\n",
119           t, VPop[0], VPop[1], VPop[2], VPop[3]);
120 }
121 return EXIT_SUCCESS;
122 }

```

123 2.2 SWIG

124 In order to use GeNN from Python, both the model creation API and the `SharedLibraryModel`
 125 functionality need to be ‘wrapped’ so they can be called from Python. While this is possible using
 126 the API built into Python itself, wrapper functions would need to be manually implemented for each GeNN
 127 function to be exposed which would result in a lot of maintenance overhead. Instead, we chose to use
 128 SWIG (Beazley, 1996) to automatically generate wrapper functions and classes. SWIG generates Python
 129 modules based on special interface files which can directly include C++ code as well as special ‘directives’
 130 which control SWIG. For example, the following SWIG interface file would wrap the C++ code in `test.h` in
 131 a Python module called `test_module` within a Python packages called `test_package`:

```

132 %module(package="test_package") test_module
133 %include "test.h"

```

134 The `%module` directive sets the name of the generated module and the package it will be located in and
 135 the `%include` directive parses and automatically generates wrapper functions for the C++ header file. We
 136 use SWIG in this manner to wrap both the model building and `SharedLibraryModel` APIs described in
 137 section 2.1. However, key parts of GeNN’s API such as the `ModelSpec::addNeuronPopulation` method
 138 employed in section 2.1, rely on C++ templates which are not directly translatable to Python. Instead, valid
 139 template instantiations need to be given a unique name in Python using the `%template` SWIG directive:

```

140 %template(addNeuronPopulationLIF) ModelSpec::addNeuronPopulation<NeuronModels::LIF>;

```

141 Having to manually add these directives whenever a model is added to GeNN would be exactly the sort
 142 of maintenance overhead we were trying to avoid by using SWIG. Therefore, when building the Python
 143 wrapper, we instead search the GeNN header files for the macros used to declare models in C++ and
 144 automatically generate SWIG `%template` directives.

145 As previously discussed, a key feature of GeNN is the ease with which it allows users to define their
 146 own neuron and synapse models as well as ‘snippets’ defining how variables and connectivity should be
 147 initialised. Beneath the syntactic sugar described in our previous work (Knight and Nowotny, 2018), new
 148 models can be defined in C++ by defining a new class derived from, for example, the `NeuronModels::Base`
 149 class. The ability to extend this system to Python was a key requirement of PyGeNN and, by using SWIG
 150 ‘director’, C++ classes can be made inheritable from Python using a single SWIG directive:

```

151 %feature("director") NeuronModels::Base;

```

152 By using these directives, GeNN can seamlessly call the Python methods used to implement new neuron
 153 and synapse models from C++.

154 **2.3 PyGeNN**

155 While GeNN *could* be used from Python via the wrapper generated using SWIG, the resultant code
 156 would be unpleasant to use directly. For example, rather than being able to specify neuron parameters
 157 using a native Python types such as lists or dictionaries, one would have to use a wrapped type such as
 158 `DoubleVector` (`[0.25, 10.0, 0.0, 0.0, 20.0, 2.0, 0.5]`). Therefore, in order to provide a more
 159 user-friendly and pythonic interface, we have built PyGeNN on top of the wrapper generated by SWIG.
 160 PyGeNN combines the separate model building and simulation stages of building a GeNN model in
 161 C++ into a single API, likely to be more familiar to users of existing Python-based model description
 162 languages such as PyNEST (Eppler et al., 2009) or PyNN (Davison et al., 2008). By combining the two
 163 stages together, PyGeNN can provide a unified dictionary-based API for initialising homogeneous and
 164 heterogeneous parameters as shown in this re-implementation of the previous example:

```

165 from pygenn import genn_wrapper, genn_model
166
167 model = genn_model.GeNNModel("float", "izhikevich")
168 model.dT = 0.1
169
170 izk_init = {"V": -65.0,
171            "U": -20.0,
172            "a": [ 0.02, 0.1, 0.02, 0.02],
173            "b": [ 0.2, 0.2, 0.2, 0.2],
174            "c": [-65.0, -65.0, -50.0, -55.0],
175            "d": [ 8.0, 2.0, 2.0, 4.0]}
176
177 pop = model.add_neuron_population("Pop", 4, "IzhikevichVariable",
178                                 {}, izk_init)
179 model.add_current_source("CS", "DC", "Pop",
180                         {"amp": 10.0}, {})
181
182 model.build()
183 model.load()
184
185 v = pop.vars["V"].view
186 while model.t < 200.0:
187     model.step_time()
188     model.pull_state_from_device("Pop")
189     print("%f, %f, %f, %f, %f"
190           % (model.t, v[0], v[1], v[2], v[3]))

```

191 Initialisation of variables with homogeneous values – such as the neurons’ membrane potential – is
 192 performed by initialisation kernels generated by GeNN and initialisation of variables with heterogeneous
 193 values – such as the `a`, `b` and `c` parameters – are copied to the GPU by PyGeNN after the model is loaded.
 194 While the PyGeNN API is more pythonic and, hopefully, more user-friendly than the C++ interface, it
 195 still provides users with the same low-level control over the simulation. Furthermore, by using SWIG’s
 196 numpy (Van Der Walt et al., 2011) interface, the host memory allocated by GeNN can be accessed directly
 197 from Python using the `pop.vars["V"].view` syntax meaning that no potentially expensive additional
 198 copying of data is required.

As illustrated in the previously-defined model, for convenience, PyGeNN allows users to access GeNN's built-in models. However, one of PyGeNN's most powerful features is that it enables users to easily define their own neuron and synapse models from within Python. For example, an Izhikevich neuron model (Izhikevich, 2003) can be defined using the `create_custom_neuron_class` helper function which provides some syntactic sugar over directly inheriting from the SWIG director class:

```

204 izk_model = genn_model.create_custom_neuron_class(
205     "izk",
206     param_names=["a", "b", "c", "d"],
207     var_name_types=[("V", "scalar"), ("U", "scalar")],
208     sim_code=
209         """
210         $(V) += 0.5 * (0.04 * $(V) * $(V) + 5.0 * $(V) + 140.0 - $(U) + $(Isyn)) * DT;
211         $(V) += 0.5 * (0.04 * $(V) * $(V) + 5.0 * $(V) + 140.0 - $(U) + $(Isyn)) * DT;
212         $(U) += $(a) * ($(b) * $(V) - $(U)) * DT;
213         """,
214     threshold_condition_code="$ (V) >= 30.0",
215     reset_code=
216         """
217         $(V) = $(c);
218         $(U) += $(d);
219         """)

```

The `param_names` list defines the real-valued parameters that are constant across the whole population of neurons and the `var_name_types` list defines the model state variables and their type (the `scalar` type is an alias for either single or double-precision floating point, depending on the precision passed to the `GeNNModel` constructor). The behaviour of the model is then defined using a number of code strings. Unlike in tools like Brian 2 (Stimberg et al., 2019), these code strings are specified in a C-like language rather than using differential equations. This language provides standard C control flow statements as well as the transcendental functions from the standard maths library. Additionally, variables provided by GeNN such as the membrane voltage in the model above can be accessed using the `$(V)` syntax and functions provided by GeNN can be called using the `$(F, 1, 2)` syntax (where `F` is a 2 argument function). Using C-like code strings allows expert users to choose their own solver for models described in terms of differential equations and to programatically define models such as spike sources. For example, in the model presented above, we chose to implement the neuron using the idiosyncratic forward Euler integration scheme employed by Izhikevich (2003). Finally, the `threshold_condition_code` expression defines *when* the neuron will spike whereas the `reset_code` code string defines how the state variables should be reset after a spike.

2.4 Spike recording system

Internally, GeNN stores the spikes emitted by a neuron population during one simulation timestep in an array containing the indices of the neurons that spiked alongside a counter of how many spikes have been emitted overall. Previously, recording spikes in GeNN was very similar to the recording of voltages shown in the previous example code – the array of neuron indices was simply copied from the GPU to the CPU every timestep. However, especially when simulating models with a small simulation timestep, such frequent synchronization between the CPU and GPU is costly – especially if a slower, interpreted language such as Python is involved. Furthermore, biological neurons typically spike at a low rate (in the cortex, the average firing rate is only around 3 Hz (Buzsáki and Mizuseki, 2014)) meaning that the amount of spike

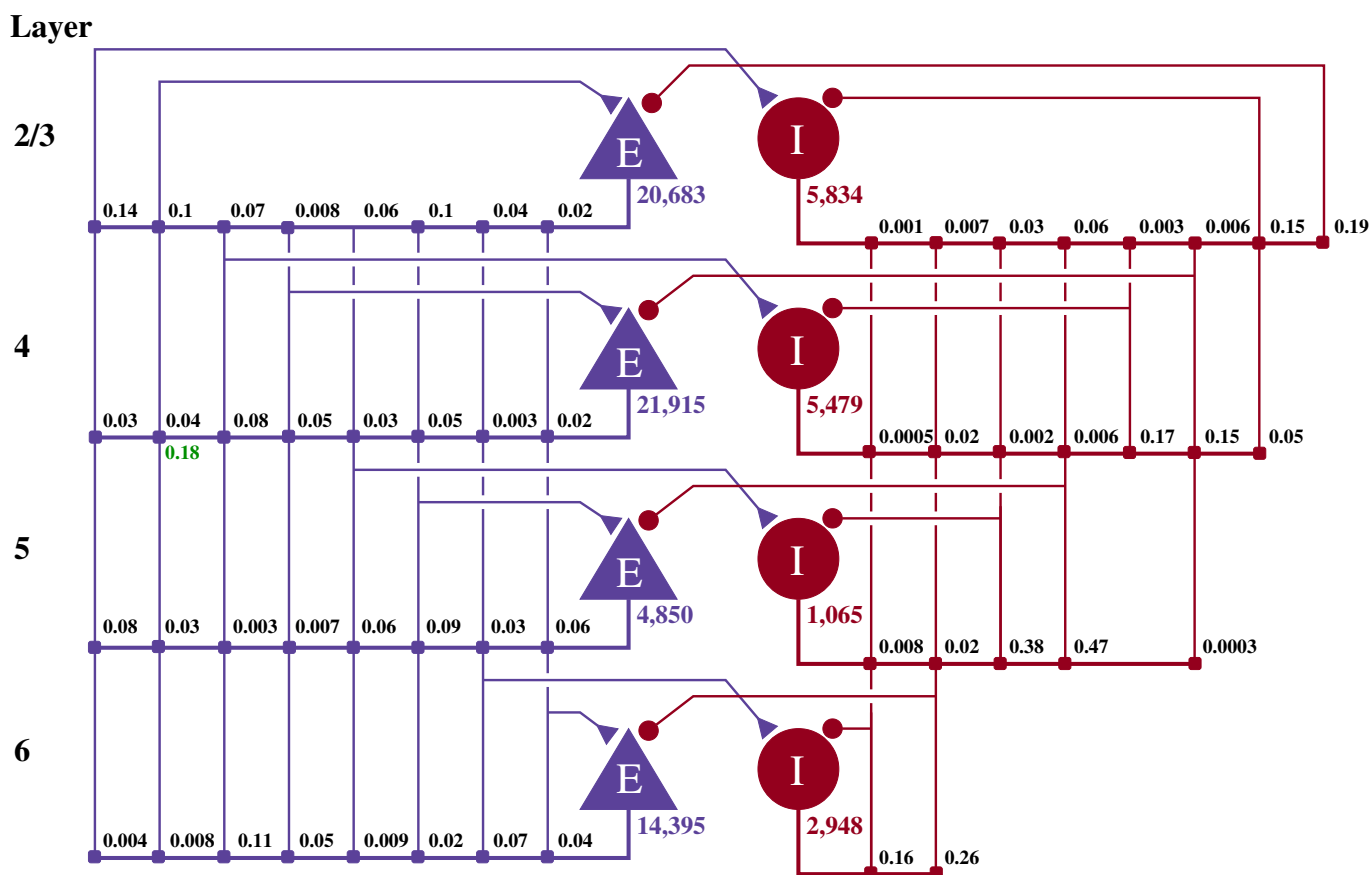


Figure 1. Illustration of the microcircuit model. Blue triangles represent excitatory populations, red circles represent inhibitory populations and the number beneath each symbol shows the number of neurons in each population. Connection probabilities are shown in small bold numbers at the appropriate point in the connection matrix. All excitatory synaptic weights are normally distributed with a mean of 0.0878 nA (unless otherwise indicated in green) and a standard deviation of 0.00878 nA. All inhibitory synaptic weights are normally distributed with a mean of 0.3512 nA and a standard deviation of 0.03512 nA.

243 data transferred every timestep is typically very small. One solution to these inefficiencies is to store many
 244 timesteps worth of spike data on the GPU and use more infrequent, larger transfers to copy it to the CPU.

245 When a model includes delays, the array of indices and the counter used to store spikes internally are
 246 duplicated for each delay ‘slot’. Additional delay slots could be artificially added to the neuron population
 247 so that this data structure could be re-used to also store spike data for subsequent recording. However,
 248 the array containing the indices has memory allocated for all neurons to handle the worst case where all
 249 neurons in the population fire in the same time step. Therefore, while this data structure is ideal for efficient
 250 spike propagation, using it to store many timesteps worth of spikes would be very wasteful of memory.
 251 At low firing rates, the most memory efficient solution would be to simply store the indices of neurons
 252 which spiked each timestep, for example in a data structure similar to a Yale sparse matrix with each ‘row’
 253 representing a timestep (Eisenstat et al., 1977). However, not only would the efficiency of this approach
 254 rely on GeNN *only* being used for models with biologically-plausible firing rates, but the amount of
 255 memory required to store the spikes for a given number of timesteps could not be determined ahead of time.
 256 Therefore, either GeNN or the user would need to regularly check the level of usage to determine whether
 257 the buffer was exhausted, leading to exactly the type of host-synchronisation overheads the spike recording
 258 system is designed to alleviate. Instead, we represent the spikes emitted by a population of N neurons in a

single simulation timestep as a N bit bitfield where a ‘1’ represents a spike and a ‘0’ the absence of one. Spiking data over multiple timesteps is then represented by a circular buffer of these bitfields. While at very low firing rates, this approach uses more memory than storing the indices of the neurons which spiked, it still allows the spiking output of relatively large models, running for many timesteps to be stored in a small amount of memory. For example, the spiking output of a model with 100×10^3 neurons running for 10×10^3 simulation timesteps, required less than 120 MB – a small fraction of the memory on a modern GPU. While efficiently handling spikes stored in a bitfield is a little trickier than working with a list of neuron indices, GeNN provides an efficient C++ helper function for saving the spikes stored in a bitfield to a text file and a numpy-based method for decoding them in PyGeNN.

2.5 Cortical microcircuit model

Potjans and Diesmann (2014) developed the cortical microcircuit model of 1 mm^3 of early-sensory cortex illustrated in figure 1. The model consists of 77 169 LIF neurons, divided into separate populations representing the excitatory and inhibitory population in each of 4 cortical layers (2/3, 4, 5 and 6). The membrane voltage V_i of each neuron i is modelled as:

$$\tau_m \frac{dV_i}{dt} = (V_{\text{rest}} - V_i) + R_m(I_{\text{syn}_i} + I_{\text{ext}_i}), \quad (1)$$

where $\tau_m = 10 \text{ ms}$ and $R_m = 40 \text{ M}\Omega$ represent the time constant and resistance of the neuron’s cell membrane, $V_{\text{rest}} = -65 \text{ mV}$ defines the resting potential, I_{syn_i} represents the synaptic input current and I_{ext_i} represents an external input current. When the membrane voltage crosses a threshold $V_{\text{th}} = -50 \text{ mV}$ a spike is emitted, the membrane voltage is reset to V_{rest} and updating of V is suspended for a refractory period $\tau_{\text{ref}} = 2 \text{ ms}$. Neurons in each population are connected randomly with numbers of synapses derived from an extensive review of the anatomical literature. These synapses are current-based, i.e. presynaptic spikes lead to exponentially-decaying input currents I_{syn_i}

$$\tau_{\text{syn}} \frac{dI_{\text{syn}_i}}{dt} = -I_{\text{syn}_i} + \sum_{j=0}^n w_{ij} \sum_{t_j} \delta(t - t_j), \quad (2)$$

where $\tau_{\text{syn}} = 0.5 \text{ ms}$ represents the synaptic time constant, w_{ij} represents the synaptic weight and t_j are the arrival times of incoming spikes from n presynaptic neurons. Within each synaptic projection, all synaptic strengths and transmission delays are normally distributed using the parameters presented in Potjans and Diesmann (2014, table 5) and, in total, the model has approximately 0.3×10^9 synapses. As well as receiving synaptic input, each neuron in the network also receives an independent Poisson input current, representing input from neighbouring not explicitly modelled cortical regions. The Poisson input is delivered to each neuron via I_{ext_i} with

$$\tau_{\text{syn}} \frac{dI_{\text{ext}_i}}{dt} = -I_{\text{ext}_i} + w_{\text{ext}} \text{Poisson}(\nu_{\text{ext}} \Delta t), \quad (3)$$

where ν_{ext} represents the mean input rate and w_{ext} represents the weight. The ordinary differential Eq. 1, 2 and 3 are solved with an exponential Euler algorithm. For a full description of the model parameters, please refer to Potjans and Diesmann (2014, tables 4 and 5) and for a description of the strategies used by GeNN to parallelise the initialisation and subsequent simulation of this network, please refer to Knight and Nowotny (2018, section 2.3). This model requires simulation using a relatively small

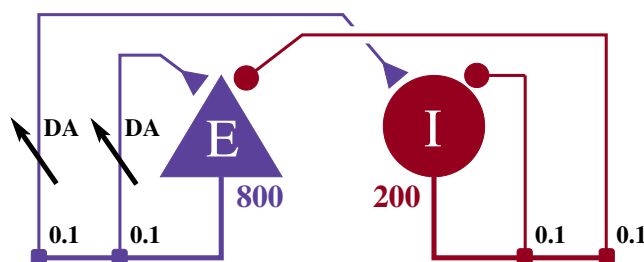


Figure 2. Illustration of the balanced random network model. The blue triangle represents the excitatory population, the red circle represents the inhibitory population, and the numbers beneath each symbol show the number of neurons in each population. Connection probabilities are shown in small bold numbers at the appropriate point in the connection matrix. All excitatory synaptic weights are plastic and initialised to 1 and all inhibitory synaptic weights are initialised to -1 .

274 timestep of 0.1 ms, making the overheads of copying spikes from the GPU every timestep particularly
275 problematic.

276 2.6 Pavlovian conditioning model

277 The cortical microcircuit model described in the previous section is ideal for exploring the performance
278 of short simulations of relatively large models. However, the performance of longer simulations of smaller
279 models is equally vital. Such models can be particularly troublesome for GPU simulation as, not only might
280 they not offer enough parallelism to fully occupy the device but, each timestep can be simulated so quickly
281 that the overheads of launching kernels etc can dominate. Additional overheads can be incurred when
282 models require injecting external stimuli throughout the simulation. Longer simulations are particularly
283 useful when exploring synaptic plasticity so, to explore the performance of PyGeNN in this scenario, we
284 simulate a model of Pavlovian conditioning using a three-factor Spike-Timing-Dependent Plasticity (STDP)
285 learning rule (Izhikevich, 2007).

286 2.6.1 Neuron model

The model illustrated in figure 2 consists of an 800 neuron excitatory population and a 200 neuron inhibitory population, within which, each neuron i is modelled using the Izhikevich model (Izhikevich, 2003) whose dimensionless membrane voltage V_i and adaption variables U_i evolve such that:

$$\frac{dV_i}{dt} = 0.04V_i^2 + 5V_i + 140 - U_i + I_{\text{syn}_i} + I_{\text{ext}_i} \quad (4)$$

$$\frac{dU_i}{dt} = a(bV_i - U_i) \quad (5)$$

287 When the membrane voltage rises above 30, a spike is emitted and V_i is reset to c and d is added to
288 U_i . Excitatory neurons use the regular-spiking parameters (Izhikevich, 2003) where $a = 0.02$, $b = 0.2$,
289 $c = -65.0$, $d = 8.0$ and inhibitory neurons use the fast-spiking parameters (Izhikevich, 2003) where
290 $a = 0.1$, $b = 0.2$, $c = -65.0$, $d = 2.0$. Again, I_{syn_i} represents the synaptic input current and I_{ext_i} represents
291 an external input current. While there are numerous ways to solve Eq. 4 and 5 (Humphries and Gurney,
292 2007; Hopkins and Furber, 2015; Pauli et al., 2018), we chose to use the idiosyncratic forward Euler
293 integration scheme employed by Izhikevich (2003) in the original work (Izhikevich, 2007). Under this
294 scheme, Eq. 4 is first integrated for two 0.5 ms timesteps and then, based on the updated value of V_i , Eq. 5
295 is integrated for a single 1 ms timestep.

296 2.6.2 Synapse models

The excitatory and inhibitory neural populations are connected recurrently, as shown in figure 1, with instantaneous current-based synapses:

$$I_{\text{syn}_i}(t) = \sum_{j=0}^n w_{ij} \sum_{t_j} \delta(t - t_j), \quad (6)$$

where t_j are the arrival times of incoming spikes from n presynaptic neurons. Inhibitory synapses are static with $w_{ij} = -1.0$ and excitatory synapses are plastic. Each plastic synapse has an eligibility trace C_{ij} as well as a synaptic weight w_{ij} and these evolve according to a three-factor STDP learning rule (Izhikevich, 2007):

$$\frac{dC_{ij}}{dt} = -\frac{C_{ij}}{\tau_c} + \text{STDP}(\Delta t)\delta(t - t_{\text{pre/post}}) \quad (7)$$

$$\frac{dw_{ij}}{dt} = -C_{ij}D_j \quad (8)$$

where $\tau_c = 1000$ ms represents the decay time constant of the eligibility trace and $\text{STDP}(\Delta t)$ describes the magnitude of changes made to the eligibility trace in response to the relative timing of a pair of pre and postsynaptic spikes with temporal difference $\Delta t = t_{\text{post}} - t_{\text{pre}}$. These changes are only applied to the trace at the times of pre and postsynaptic spikes as indicated by the Dirac delta function $\delta(t - t_{\text{pre/post}})$. Here, a double exponential STDP kernel is employed such that:

$$\text{STDP}(\Delta t) = \begin{cases} A_+ \exp\left(-\frac{\Delta t}{\tau_+}\right) & \text{if } \Delta t > 0 \\ A_- \exp\left(\frac{\Delta t}{\tau_-}\right) & \text{if } \Delta t < 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where the time constant of the STDP window $\tau_+ = \tau_- = 20$ ms and the strength of potentiation and depression are $A_+ = 0.1$ and $A_- = 0.15$ respectively. Finally, each excitatory neuron has an additional variable D_j which describes extracellular dopamine concentration:

$$\frac{D_j}{t} = -\frac{D_j}{\tau_d} + \text{DA}(t) \quad (10)$$

297 where $\tau_d = 200$ ms represents the time constant of dopamine uptake and $\text{DA}(t)$ the dopamine input over
298 time.

299 2.6.3 PyGeNN implementation of three-factor STDP

300 The first step in implementing this learning rule in PyGeNN is to implement the STDP updates and decay
301 of C_{ij} using GeNN's event-driven plasticity system, the implementation of which was described in our
302 previous work (Knight and Nowotny, 2018). Using a similar syntax to that described in section 2.3, we first
303 create a new 'weight update model' with the learning rule parameters and the w_{ij} and C_{ij} state variables:

```
304 izhikevich_stdp_model = create_custom_weight_update_class(  
305     "izhikevich_stdp",  
306
```

```

307     param_names=["tauPlus", "tauMinus",
308                 "tauC", "aPlus", "aMinus"],
309     var_name_types=[("w", "scalar"), ("c", "scalar")],

```

310 We then instruct GeNN to record the times of current and previous pre and postsynaptic spikes. The current
 311 spike time will equal the current time if a spike of this sort is being processed in the current timestep
 312 whereas the previous spike time only tracks spikes which have occurred *before* the current timestep:

```

313     is_pre_spike_time_required=True,
314     is_post_spike_time_required=True,
315
316     is_prev_pre_spike_time_required=True,
317     is_prev_post_spike_time_required=True,

```

318 Next we define the ‘sim code’ which is called whenever presynaptic spikes arrive at the synapse. This code
 319 first implements Eq. 6 – adding the synaptic weight (w_{ij}) to the postsynaptic neuron’s input (I_{syn_i}) using
 320 the `$(addToInSyn, x)` function.

```

321     sim_code=
322         """
323         $(addToInSyn, $(w));

```

324 Within the sim code we also need to calculate the time that has elapsed since the last update of C_{ij} using
 325 the spike times we previously requested that GeNN record. Within a timestep, GeNN processes presynaptic
 326 spikes before postsynaptic spikes so the time of the last update to C_{ij} will be the latest time either type of
 327 spike was processed in previous timesteps:

```

328         const scalar tc = fmax($(prev_sT_pre),
329                                $(prev_sT_post));

```

330 Using this time, we can now calculate how much to decay C_{ij} using the closed-form solution to Eq. 7:

```

331         const scalar tagDecay = exp(-($(t) - tc) / $(tauC));
332         scalar newTag = $(c) * tagDecay;

```

333 To complete the sim code we calculate the depression case of Eq. 9 (here we use the *current* postsynaptic
 334 spike time as, if a postsynaptic and presynaptic spike occur in the same timestep, there should be no
 335 update).

```

336         const scalar dt = $(t) - $(sT_post);
337         if (dt > 0) {
338             newTag -= ($(aMinus) * exp(-dt / $(tauMinus)));
339         }
340         $(c) = newTag;
341         """ ,

```

342 Finally we define the ‘learn post code’ which is called whenever a postsynaptic spike arrives at the synapse.
 343 Other than implementing the potentiation case of Eq. 9 and using the *current* presynaptic spike time when
 344 calculating the time since the last update of C_{ij} – in order to correctly handle presynaptic updates made in
 345 the same timestep – this code is very similar to the sim code:

```

346 learn_post_code=
347     """
348     const scalar tc = fmax($(sT_pre),
349                             $(prev_sT_post));
350
351     const scalar tagDecay = exp(-$(t) - tc) / $(tauC));
352     scalar newTag = $(c) * tagDecay;
353
354     const scalar dt = $(t) - $(sT_pre);
355     if (dt > 0) {
356         newTag += ($(aPlus) * exp(-dt / $(tauPlus)));
357     }
358     $(c) = newTag;
359     """

```

360 Adding the synaptic weight w_{ij} update described by Eq. 8 requires two further additions to the model.
361 As well as the pre and postsynaptic spikes, the weight update model needs to receive events whenever
362 dopamine is injected via DA. GeNN supports such events via the ‘spike-like event’ system which allows
363 events to be triggered based on an expression evaluated on the presynaptic neuron. In this case, this
364 expression simply tests an `injectDopamine` flag which gets set by the dopamine injection logic in our
365 presynaptic neuron model:

```

366 event_threshold_condition_code="injectDopamine",

```

367 In order to extend our event-driven update of C_{ij} to include spike-like events we need to instruct GeNN to
368 record the times at which they occur:

```

369 is_pre_spike_event_time_required=True,
370 is_prev_pre_spike_event_time_required=True,

```

371 The spike-like events can now be handled using a final ‘event code’ string:

```

372 event_code=
373     """
374     const scalar tc = fmax($(sT_pre), fmax($(prev_sT_post), $(prev_seT_pre)));
375     const scalar tagDecay = exp(-$(t) - tc) / $(tauC));
376     $(c) *= tagDecay;
377     """

```

After updating the previously defined calculations of `tc` in the sim code and learn post code in the same way to also include the times of spike-like events, all that remains is to update w_{ij} . Mikaitis et al. (2018) showed how Eq. 8 could be solved algebraically, allowing w_{ij} to be updated in an event-driven manner with:

$$\Delta w_{ij} = \frac{C(t_c^{last})D(t_d^{last})}{-\left(\frac{1}{\tau_c} + \frac{1}{\tau_d}\right)} \left(e^{-\frac{t-t_c^{last}}{\tau_c}} e^{-\frac{t-t_d^{last}}{\tau_d}} - e^{-\frac{t_w^{last}-t_c^{last}}{\tau_c}} e^{-\frac{t_w^{last}-t_d^{last}}{\tau_d}} \right) \quad (11)$$

where t_c^{last} , t_w^{last} and t_d^{last} represent the last times at which C_{ij} , W_{ij} and D_j respectively were updated. Because we will always update w_{ij} and C_{ij} together when presynaptic, postsynaptic and spike-like events

occur, $t_c^{last} = t_w^{last}$ and Eq. 11 can be simplified to:

$$\Delta w_{ij} = \frac{C(t_c^{last})D(t_d^{last})}{-\left(\frac{1}{\tau_c} + \frac{1}{\tau_d}\right)} \left(e^{-\frac{t-t_c^{last}}{\tau_c}} e^{-\frac{t-t_d^{last}}{\tau_d}} - e^{-\frac{t_c^{last}-t_d^{last}}{\tau_d}} \right) \quad (12)$$

and this update can now be added to each of our three event handling code strings to complete the implementation of the learning rule.

2.6.4 PyGeNN implementation of Pavlovian conditioning experiment

To perform the Pavlovian conditioning experiment described by Izhikevich (2007) using this model, we chose 100 random groups of 50 neurons (each representing stimuli $S_1 \dots S_{100}$) from amongst the two neural populations. Stimuli are presented to the network in a random order, separated by intervals sampled from $U(100, 300)$ ms. The neurons associated with an active stimulus are stimulated for a single 1 ms simulation timestep with a current of 40.0 nA, in addition to the random background current of $U(-6.5, 6.5)$ nA, delivered to each neuron via I_{ext_i} throughout the simulation. S_1 is arbitrarily chosen as the Conditional Stimuli (CS) and, whenever this stimuli is presented, a reward in the form of an increase in dopamine is delivered by setting $DA(t) = 0.5$ after a delay sampled from $U(0, 1000)$ ms. This delay period is large enough to allow a few irrelevant stimuli to be presented which act as distractors. The simplest way to implement this stimulation regime is to add a current source to the excitatory and inhibitory neuron populations which adds the uniformly-distributed input current to an externally-controllable per-neuron current. In PyGeNN, the following model can be defined to do just that:

```
stim_noise_model = create_custom_current_source_class(
    "stim_noise",
    param_names=["n"],
    var_name_types=[("iExt", "scalar", VarAccess_READ_ONLY)],
    injection_code=
        """
        $(injectCurrent, $(iExt) + ($(gennrand_uniform) * $(n) * 2.0) - $(n));
        """
```

where the `n` parameter sets the magnitude of the background noise, the `$(injectCurrent, I)` function injects a current of I nA into the neuron and `$(gennrand_uniform)` samples from $U(0, 1)$ using the ‘XORWOW’ pseudo-random number generator provided by cuRAND (NVIDIA Corporation, 2019). Once a current source population using this model has been instantiated and a memory view to `iExt` obtained in the manner described in section 2.3, in timesteps when stimulus injection is required, current can be injected into the list of neurons contained in `stimuli_input_set` with:

```
curr_ext_view[stimuli_input_set] = 40.0
curr_pop.push_var_to_device("iExt")
```

The same approach can then be used to zero the current afterwards.

3 RESULTS

In the following subsections we will analyse the performance of the models introduced in sections 2.5 and 2.6 on a representative selection of NVIDIA GPU hardware:

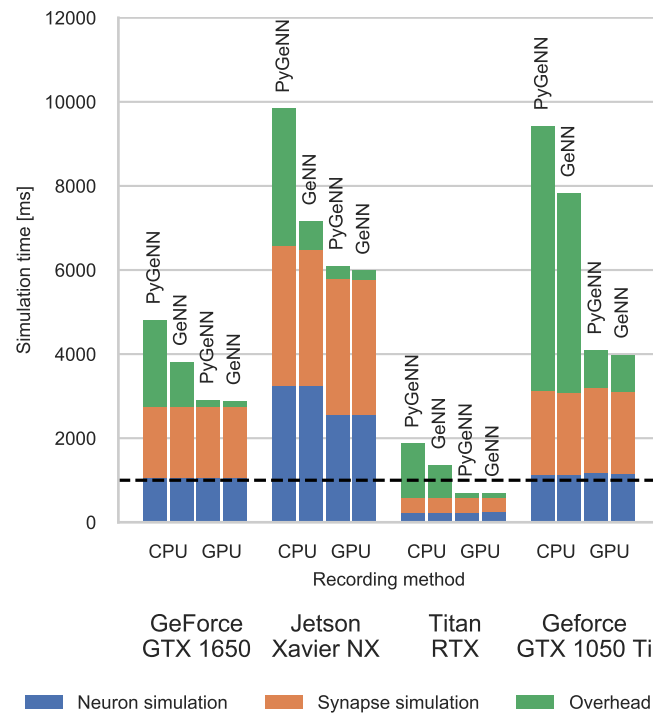


Figure 3. Simulation times of the microcircuit model running on various GPU hardware for 1 s of biological time. ‘Overhead’ refers to time spent in simulation loop but not within CUDA kernels. The dashed horizontal line indicates realtime performance

- Jetson Xavier NX – a low-power embedded system with a GPU based on the Volta architecture with 8 GB of shared memory.
- GeForce GTX 1050Ti – a low-end desktop GPU based on the Pascal architecture with 4 GB of dedicated memory.
- GeForce GTX 1650 – a low-end desktop GPU based on the Turing architecture with 4 GB of dedicated memory.
- Titan RTX – a high-end workstation GPU based on the Turing architecture with 24 GB of dedicated memory.

All of these systems run Ubuntu 18 apart from the system with the GeForce 1050 Ti which runs Windows 10.

3.1 Cortical microcircuit model performance

Figure 3 shows the simulation times for the full-scale microcircuit model. We measured the total simulation time by querying the `std::chrono::high_resolution_clock` in C++ and the `time.perf_counter` in Python before and after the simulation loop; and used CUDA’s own event timing system (NVIDIA Corporation, 2021, Section 3.2.6.7) to record the time taken by the neuron and synapse kernels. As one might predict, the Jetson Xavier NX is slower than the three desktop GPUs but, considering that it only consumes a maximum of 15 W compared to 75 W or 320 W for the GeForce cards and Titan RTX respectively, it still performs impressively. The time taken to actually simulate the models (‘Neuron simulation’ and ‘Synapse simulation’) are the same when using PyGeNN and GeNN as all optimisation options are exposed to PyGeNN. Interestingly, when simulating *this* model, the larger L1 cache and

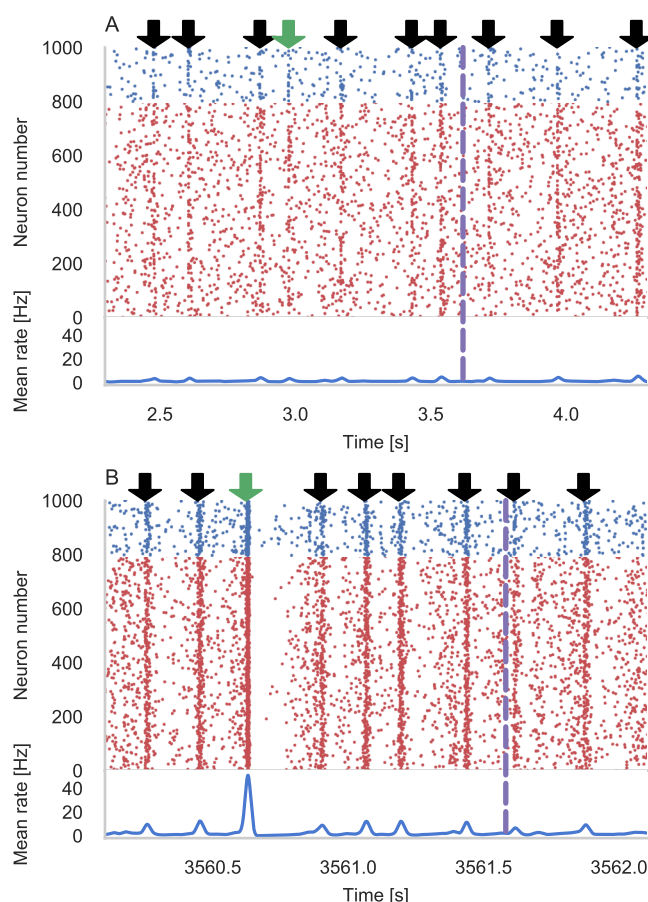


Figure 4. Results of Pavlovian conditioning experiment. Raster plot and spike density function (SDF) (Szücs, 1998) showing activity centred around first delivery of Conditional Stimulus (CS) during initial (A) and final (B) 50 s of simulation. Downward green arrows indicate times at which CS is delivered and downward black arrows indicate times when other, un-rewarded stimuli are delivered. Vertical dashed lines indicate times at which dopamine is delivered. The population SDF was calculated by convolving the spikes with a Gaussian kernel of $\sigma = 10$ ms width.

architectural improvements present in the Turing-based GTX 1650 do not result in significantly improved performance over the Pascal-based GTX 1050Ti. Instead, the slightly improved performance of the GTX 1650 can probably be explained by its additional 128 CUDA cores.

Without the recording system described in section 2.4, the CPU and GPU need to be synchronised after every timestep to allow spike data to be copied off the GPU and stored in a suitable data structure. The ‘overheads’ shown in figure 3 indicate the time taken by these processes as well as the unavoidable overheads of launching CUDA kernels etc. Because Python is an interpreted language, updating the spike data structures is somewhat slower and this is particularly noticeable on devices with a slower CPU such as the Jetson Xavier NX. However, unlike the desktop GPUs, the Jetson Xavier NX’s 8 GB of memory is shared between the GPU and the CPU meaning that data does not need to be copied between their memories and can instead be accessed by both. While, using this shared memory for recording spikes reduces the overhead of copying data off the device, because the GPU and CPU caches are not coherent, caching must be disabled on this memory which reduces the performance of the neuron kernel. Although the Windows machine has a relatively powerful CPU, the overheads measured in both the PyGeNN and GeNN simulations run on this system are extremely large due to additional queuing between the application and

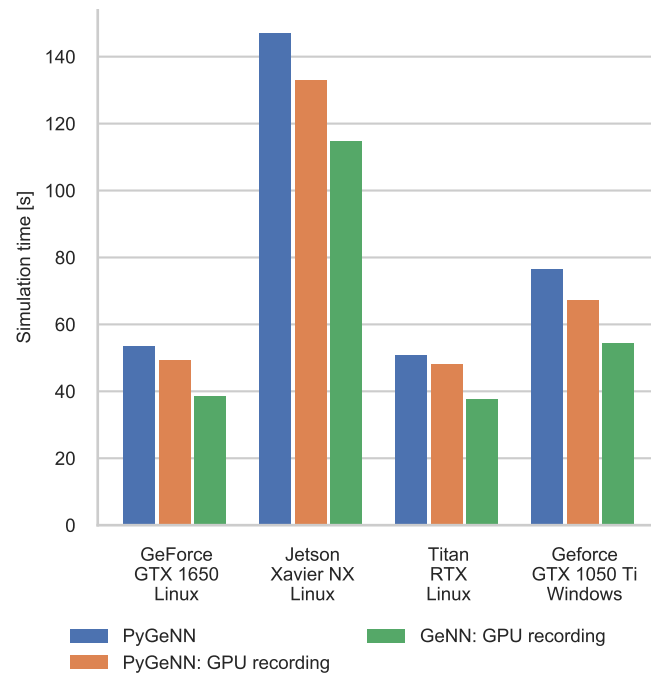


Figure 5. Simulation times of the Pavlovian Conditioning model running on various GPU hardware for 1 h of biological time. ‘GPU recording’ indicates simulations where the new recording system is employed. Times are taken from averages calculated over 5 runs of each model.

the GPU driver caused by the Windows Display Driver Model (WDDM). When small – in this case 0.1 ms – simulation timesteps are used, this makes per-timestep synchronisation disproportionately expensive.

However, when the spike recording system described in section 2.4 is used, spike data is kept in GPU memory until the end of the simulation and overheads are reduced by up to 10×. Because synchronisation with the CPU is no longer required every timestep, simulations run approximately twice as fast on the Windows machine. Furthermore, on the high-end desktop GPU, the simulation now runs faster than real-time in both PyGeNN and GeNN versions – significantly faster than other recently published GPU simulators (Golosio et al., 2021) and even specialised neuromorphic systems (Rhodes et al., 2020).

3.2 Pavlovian conditioning performance

Figure 4 shows the results of an example simulation of the Pavlovian conditioning model. At the beginning of each simulation (Figure 4A), the neurons representing every stimulus respond equally. However, after 1 h of simulation, the response to the CS becomes much stronger (Figure 4B) – showing that these neurons have been selectively associated with the stimulus even in the presence of the distractors and the delayed reward. In figure 5, we show the runtime performance of simulations of the Pavlovian conditioning model, running on the GPUs described above using PyGeNN with and without the recording system described in section 2.4. These PyGeNN results are compared to a GeNN simulation which also uses the recording system. Because each simulation timestep only takes a few μ s, the overhead of using CUDA timing events significantly alters the performance so, for this model, we only measure the duration of the simulation loop using the approaches described in the previous section. Although we only record the spiking activity during the first and last 50 s, using the recording system still significantly improves the overall performance on all devices – especially on the Jetson Xavier NX with its slower CPU. Interestingly the Titan RTX and GTX 1650 perform identically in this benchmark with speedups ranging from 62× to 72× real-time. This

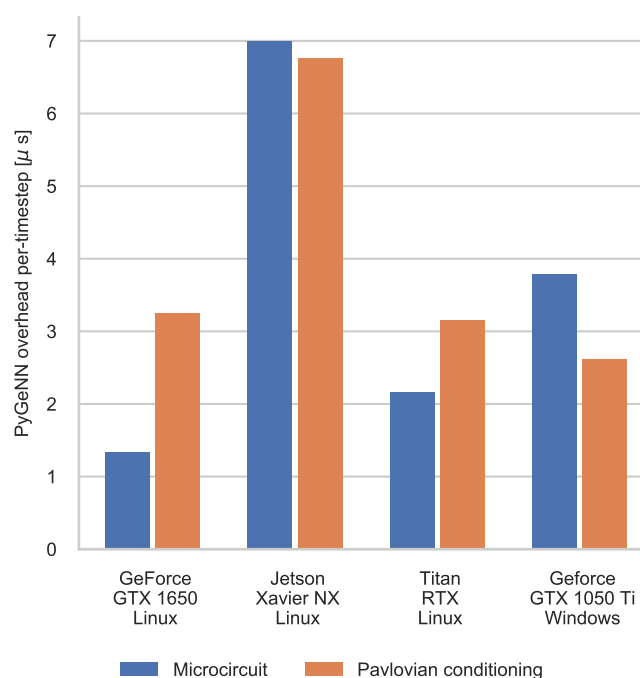


Figure 6. Comparison of the duration of individual timestep in PyGeNN and GeNN simulation in microcircuit and Pavlovian conditioning experiments. Times are taken from averages calculated over 5 runs using the GPU recording system.

is because, as discussed previously, this model is simply not large enough to fill the 4608 CUDA cores present on the Titan RTX. Therefore, as the two GPUs share the same Turing architecture and have very similar clock speeds (1350 MHz–1770 MHz for the Titan RTX and 1485 MHz–1665 MHz for the GTX 1650), the two GPUs perform very similarly. As for the simulations of the microcircuit model, the Jetson Xavier NX performs rather slower than the desktop GPUs but still achieves speedups of up to $31\times$.

Interestingly, unlike in the simulations of the microcircuit model, here the GTX 1050 Ti performs rather differently. Although the clock speed of this device is approximately the same as the other GPUs (1290 MHz–1392 MHz) and it has a similar number of CUDA cores to the GTX 1650, its performance is significantly worse. The difference in performance across all configurations is likely to be due to architectural differences between the older Pascal; and newer Volta and Turing architectures. Specifically, Pascal GPUs have one type of Arithmetic Logic Unit (ALU) which handles both integer and floating point arithmetic whereas, the newer Volta and Turing architectures have equal numbers of dedicated integer and floating point ALUs as well as significantly larger L1 caches. As discussed in our previous work (Knight and Nowotny, 2018), these architectural features are particularly beneficial for SNN simulations with STDP where a large amount of floating point computation is required to update the synaptic state *and* additional integer arithmetic is required to calculate the indices into the sparse matrix data structures.

The difference between the speeds of the PyGeNN and GeNN simulations of the Pavlovian conditioning model (figure 5) *appear* much larger than those of the microcircuit model (figure 3). However, as figure 6 illustrates, for individual timesteps the excess time due to overheads is approximately the same for both models and consistent with the cost of a small number of Python to C++ function calls (Apache Crail, 2019). Depending on the size and complexity of the model as well as the hardware used, this overhead may or may not be important. For example, when simulating the microcircuit model for 1 s on the Titan RTX,

the overhead of using PyGeNN is less than 0.2 % but, when simulating the Pavlovian conditioning model on the same device, the overhead of using PyGeNN is almost 31 %.

4 DISCUSSION

In this paper we have introduced PyGeNN, a Python interface to the C++ based GeNN library for GPU accelerated spiking neural network simulations.

Uniquely, the new interface provides access to all the features of GeNN, without leaving the comparative simplicity of Python and with, as we have shown, typically negligible overheads from the Python bindings. PyGeNN also allows bespoke neuron and synapse models to be defined from within Python, making PyGeNN much more flexible and broadly applicable than, for instance, the Python interface to NEST (Eppler et al., 2009) or the PyNN model description language used to expose CARLsim to Python (Balaji et al., 2020).

In many ways, the new interface resembles elements of the Python-based Brian 2 simulator (Stimberg et al., 2019) (and its Brian2GeNN backend (Stimberg et al., 2020)) with two key differences. Unlike in Brian 2, bespoke models in PyGeNN are defined with ‘C-like’ code snippets. This has the advantage of unparalleled flexibility for the expert user, but comes at the cost of more complexity as the code for a timestep update needs to include a suitable solver as well as merely differential equations. The second difference lies in how data structures are handled. Whereas simulations run using the C++ or Brian2GeNN Brian 2 backends use files to exchange data with Python, the underlying GeNN data structures are directly accessible from PyGeNN meaning that no disk access is involved.

As we have demonstrated, the PyGeNN wrapper, exactly like native GeNN, can be used on a variety of hardware from data centre scale down to mobile devices such as the NVIDIA Jetson. This allows for the same codes to be used in large-scale brain simulations and embedded and embodied spiking neural network research. Supporting the popular Python language in this interface makes this ecosystem available to a wider audience of researchers in both Computational Neuroscience, bio-mimetic machine learning and autonomous robotics.

The new interface also opens up opportunities to support researchers that work with other Python based systems. In the Computational Neuroscience and Neuromorphic computing communities, we can now build a PyNN (Davison et al., 2008) interface on top of PyGeNN and, in fact, a prototype of such an interface is in development. Furthermore, for the burgeoning spike-based machine learning community, we can use PyGeNN as the basis for a spike-based machine learning framework akin to TensorFlow or PyTorch for rate-based models. A prototype interface of this sort called mlGeNN is in development and close to release.

In this work we have introduced a new spike recording system for GeNN and have shown that, using this system, we can now simulate the Potjans microcircuit model (Potjans and Diesmann, 2014) faster than real-time and, to the best of our knowledge, faster than any other system. Finally, the excellent performance we have demonstrated using low-end Turing architecture GPUs is very exciting in terms of increasing the accessibility of GPU accelerated Computational Neuroscience and SNN machine learning research.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

JK and TN wrote the paper. TN is the original developer of GeNN. AK was the original developer of PyGeNN. JK is currently the primary developer of both GeNN and PyGeNN and was responsible for implementing the spike recording system. JK performed the experiments and the analysis of the results that are presented in this work.

FUNDING

This work was funded by the EPSRC (Brains on Board project, grant number EP/P006094/1 and ActiveAI project, grant number EP/S030964/1), the European Union's Horizon 2020 research and innovation program under Grant Agreement 945539 (HBP SGA3) and a Google Summer of Code grant to AK.

ACKNOWLEDGMENTS

We would like to thank Malin Sandström and everyone else at the International Neuroinformatics Coordinating Facility (INCF) for their hard work running the Google Summer of Code mentoring organisation every year. Without them, this and many other exciting Neuroinformatics projects would not be possible.

DATA AVAILABILITY STATEMENT

All models, data and analysis scripts used for this study can be found in https://github.com/BrainsOnBoard/pygenn_paper. All experiments were carried out using the GeNN 4.4.0 which is fully open source and available from <https://doi.org/10.5281/zenodo.4419159>.

REFERENCES

- Akar, N. A., Cumming, B., Karakasis, V., Kusters, A., Klijn, W., Peyser, A., et al. (2019). Arbor — A Morphologically-Detailed Neural Network Simulation Library for Contemporary High-Performance Computing Architectures. In *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (IEEE), 274–282. doi:10.1109/EMPDP.2019.8671560
- [Dataset] Apache Crail (2019). Crail Python API: Python -> C/C++ call overhead
- Balaji, A., Adiraju, P., Kashyap, H. J., Das, A., Krichmar, J. L., Dutt, N. D., et al. (2020). PyCARL: A PyNN Interface for Hardware-Software Co-Simulation of Spiking Neural Network
- Bautembach, D., Oikonomidis, I., and Argyros, A. (2021). Multi-GPU SNN Simulation with Perfect Static Load Balancing
- Beazley, D. M. (1996). Using SWIG to control, prototype, and debug C programs with Python. In *Proc. 4th Int. Python Conf*
- Buzsáki, G. and Mizuseki, K. (2014). The log-dynamic brain: how skewed distributions affect network operations. *Nature reviews. Neuroscience* 15, 264–78. doi:10.1038/nrn3687
- Carnevale, N. T. and Hines, M. L. (2006). *The NEURON book* (Cambridge University Press)
- Chou, T.-s., Kashyap, H. J., Xing, J., Listopad, S., Rounds, E. L., Beyeler, M., et al. (2018). CARLsim 4: An Open Source Library for Large Scale, Biologically Detailed Spiking Neural Network Simulation using Heterogeneous Clusters. In *2018 International Joint Conference on Neural Networks (IJCNN)* (IEEE), 1–8. doi:10.1109/IJCNN.2018.8489326

- 560 Davison, A. P., Brüderle, D., Eppler, J., Kremkow, J., Müller, E., Pecevski, D., et al. (2008). PyNN: A
 561 Common Interface for Neuronal Network Simulators. *Frontiers in neuroinformatics* 2, 11. doi:10.3389/
 562 neuro.11.011.2008
- 563 Eisenstat, S. C., Gursky, M., Schultz, M. H., and Sherman, A. H. (1977). *Yale sparse matrix package. i. the*
 564 *symmetric codes*. Tech. rep., YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE
- 565 Eppler, J. M., Helias, M., Müller, E., Diesmann, M., and Gewaltig, M. O. (2009). PyNEST: A convenient
 566 interface to the NEST simulator. *Frontiers in Neuroinformatics* 2, 1–12. doi:10.3389/neuro.11.012.2008
- 567 Gewaltig, M.-O. and Diesmann, M. (2007). NEST (NEural Simulation Tool). *Scholarpedia* 2, 1430
- 568 Givon, L. E. and Lazar, A. A. (2016). Neurokernel: An open source platform for emulating the fruit fly
 569 brain. *PLOS ONE* 11, 1–25. doi:10.1371/journal.pone.0146581
- 570 Golosio, B., Tiddia, G., De Luca, C., Pastorelli, E., Simula, F., and Paolucci, P. S. (2021). Fast Simulations
 571 of Highly-Connected Spiking Cortical Models Using GPUs. *Frontiers in Computational Neuroscience*
 572 15, 1–17. doi:10.3389/fncom.2021.627620
- 573 Hines, M. L., Davison, A. P., and Müller, E. (2009). NEURON and Python. *Frontiers in Neuroinformatics*
 574 3, 1–12. doi:10.3389/neuro.11.001.2009
- 575 Hopkins, M. and Furber, S. B. (2015). Accuracy and Efficiency in Fixed-Point Neural ODE Solvers.
 576 *Neural computation* 27, 2148–2182
- 577 Humphries, M. D. and Gurney, K. (2007). Solution Methods for a New Class of Simple Model Neurons M.
 578 *Neural Computation* 19, 3216–3225. doi:doi:10.1162/neco.2007.19.12.3216
- 579 Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9,
 580 90–95. doi:10.1109/MCSE.2007.55
- 581 Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks* 14,
 582 1569–72. doi:10.1109/TNN.2003.820440
- 583 Izhikevich, E. M. (2007). Solving the Distal Reward Problem through Linkage of STDP and Dopamine
 584 Signaling. *Cerebral Cortex* 17, 2443–2452. doi:10.1093/cercor/bhl152
- 585 Knight, J. C. and Nowotny, T. (2018). GPUs Outperform Current HPC and Neuromorphic Solutions
 586 in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model. *Frontiers in*
 587 *Neuroscience* 12, 1–19. doi:10.3389/fnins.2018.00941
- 588 Knight, J. C. and Nowotny, T. (2020). Larger GPU-accelerated brain simulations with procedural
 589 connectivity. *bioRxiv* doi:10.1101/2020.04.27.063693
- 590 Mikaitis, M., Pineda García, G., Knight, J. C., and Furber, S. B. (2018). Neuromodulated Synaptic
 591 Plasticity on the SpiNNaker Neuromorphic System 12, 1–13. doi:10.3389/fnins.2018.00105
- 592 Millman, K. J. and Aivazis, M. (2011). Python for scientists and engineers. *Computing in Science and*
 593 *Engineering* 13, 9–12. doi:10.1109/MCSE.2011.36
- 594 NVIDIA, Vingelmann, P., and Fitzek, F. H. (2020). *CUDA*, developer.nvidia.com/cuda-toolkit
- 595 NVIDIA Corporation (2019). *cuRAND Library*, docs.nvidia.com/cuda/pdf/CURAND_Library.pdf
- 596 NVIDIA Corporation (2021). *CUDA C Programming Guide*,
 597 docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf
- 598 Pauli, R., Weidel, P., Kunkel, S., and Morrison, A. (2018). Reproducing Polychronization: A Guide to
 599 Maximizing the Reproducibility of Spiking Network Models. *Frontiers in Neuroinformatics* 12, 1–21.
 600 doi:10.3389/fninf.2018.00046
- 601 Plotnikov, D., Blundell, I., Ippen, T., Eppler, J. M., Rumpe, B., and Morrison, A. (2016). NESTML: a
 602 modeling language for spiking neurons. In *Lecture Notes in Informatics (LNI)*. Modellierung 2016,
 603 Karlsruhe (Germany), 17 Mar 2016 - 19 Mar 2016 (Gesellschaft für Informatik e.V. (GI)), vol. P-254,
 604 93–108

- Potjans, T. C. and Diesmann, M. (2014). The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model. *Cerebral Cortex* 24, 785–806. doi:10.1093/cercor/bhs358
- Rhodes, O., Peres, L., Rowley, A. G. D., Gait, A., Plana, L. A., Brenninkmeijer, C., et al. (2020). Real-time cortical simulation on neuromorphic hardware. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378, 20190160. doi:10.1098/rsta.2019.0160
- Stimberg, M., Brette, R., and Goodman, D. F. (2019). Brian 2, an intuitive and efficient neural simulator. *eLife* 8, 1–41. doi:10.7554/eLife.47314
- Stimberg, M., Goodman, D. F., and Nowotny, T. (2020). Brian2GeNN: accelerating spiking neural network simulations with graphics hardware. *Scientific Reports* 10, 1–12. doi:10.1038/s41598-019-54957-7
- Szücs, A. (1998). Applications of the spike density function in analysis of neuronal firing patterns. *J Neurosci Methods* 81, 159–167
- Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering* 13, 22–30. doi:10.1109/MCSE.2011.37
- Vitay, J., Dinkelbach, H., and Hamker, F. (2015). ANNarchy: a code generation approach to neural simulations on parallel hardware. *Frontiers in Neuroinformatics* 9, 19. doi:10.3389/fninf.2015.00019
- Yavuz, E., Turner, J., and Nowotny, T. (2016). GeNN: a code generation framework for accelerated brain simulations. *Scientific reports* 6, 18854. doi:10.1038/srep18854