

PyGeNN: A Python library for GPU-enhanced neural networks

James C Knight^{1,*}, Anton Komissarov², Thomas Nowotny¹

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

²(TODO: ANTON'S AFFILIATION)

Correspondence*:

James C Knight

J.C.Knight@sussex.ac.uk

2 ABSTRACT

3 For full guidelines regarding your manuscript please refer to Author Guidelines.

4 As a primary goal, the abstract should render the general significance and conceptual advance
5 of the work clearly accessible to a broad readership. References should not be cited in the
6 abstract. Leave the Abstract empty if your article does not require one, please see Summary
7 Table for details according to article type.

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

1 INTRODUCTION

10 A wide range of spiking neural network (SNN) simulators are available, each with their own application
11 domains. NEST (Gewaltig and Diesmann, 2007) is widely used for large-scale point neuron simulations
12 on distributed computing systems; NEURON (Carnevale and Hines, 2006) and Arbor (Akar et al., 2019)
13 specialise in the simulation of complex multi-compartmental models; NeuroKernel (Givon and Lazar, 2016)
14 is focused on emulating fly brain circuits using Graphics Processing Units (GPUs); and CARLsim (Chou
15 et al., 2018), ANNarchy (Vitay et al., 2015), NeuronGPU (Golosio et al., 2020) and GeNN (Yavuz et al.,
16 2016) use GPUs to accelerate point neuron models. For performance reasons, many of these simulators are
17 written in C++ and, especially amongst the older simulators, users describe their models either using a
18 Domain-Specific Language (DSL) or directly in C++. For programming language purists, a DSL may be an
19 elegant way of describing an SNN network model and, for simulator developers, not having to add bindings
20 to another language is convenient. However, both choices act as a barrier to potential users. Therefore, with
21 both the computational neuroscience and machine learning communities gradually coalescing towards a
22 Python-based ecosystem with a wealth of mature libraries for scientific computing (Hunter, 2007; Van Der
23 Walt et al., 2011; Millman and Aivazis, 2011), exposing spiking neural network simulators to Python seems
24 a pragmatic choice. NEST (Eppler et al., 2009), NEURON (Hines et al., 2009) and CARLsim (Balaji et al.,
25 2020) have all taken this route and now offer a Python interface. Furthermore, newer simulators such as
26 Arbor and Brian2 (Stimberg et al., 2019) have been designed from the ground up with a Python interface.

27 While we have recently demonstrated some very competitive performance results (Knight and Nowotny,
28 2018, 2020) using our GeNN simulator (Yavuz2016), it has so far not been usable directly from Python.

GeNN can already be used as a backend for the Python-based Brian2 simulator (Stimberg et al., 2019) but, while Brian2GeNN (Stimberg et al., 2020) allows Brian2 users to harness the performance benefits GeNN provides, it is not possible to expose all of GeNN’s unique features to Python 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. While implementing new neuron and synapse models in the majority of other GPU simulators requires extending the underlying C++ code, using PyGeNN, models can be defined directly from Python. Finally, 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 code for the simulation of spiking neural network models. GeNN handles much of the complexity of using CUDA directly as well as automatically performing device-specific optimizations so as to maximize performance.

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

```
void modelDefinition(ModelSpec &model)
{
    model.setDT(0.1);
    model.setName("izhikevich");

    NeuronModels::IzhikevichVariable::VarValues popInit(
        -65.0, -20.0, uninitialisedVar(), uninitialisedVar(),
        uninitialisedVar(), uninitialisedVar());

    model.addNeuronPopulation<NeuronModels::IzhikevichVariable>(
```

```

70         "Pop", 4, {}, popInit);
71
72     model.addCurrentSource<CurrentSourceModels::DC>(
73         "CS", "Pop", {10.0}, {});
74 }

```

75 The *genn-buildmodel* command line tool is then used to compile this file; link it against the main GeNN library and the desired backend library; and finally run the resultant executable to generate the source code required to build a simulation dynamic library (a .dll file on Windows or a .so file on Linux and Mac). 76 This dynamic library can then either be statically linked against a simulation loop provided by the user or 77 dynamically loaded by the user's simulation code. To demonstrate this latter approach, this example uses 78 the *SharedLibraryModel* helper class supplied with GeNN to dynamically load the previously defined 79 model, initialise the heterogenous neuron parameters and print each neuron's membrane voltage every 80 timestep: 81

```

83 #include "sharedLibraryModel.h"
84
85 int main()
86 {
87     SharedLibraryModel<float> model("./", "izhikevich");
88     model.allocateMem();
89     model.initialize();
90     float *aPop = model.getScalar<float>("a");
91     float *bPop = model.getScalar<float>("b");
92     float *cPop = model.getScalar<float>("c");
93     float *dPop = model.getScalar<float>("d");
94     aPop[0] = 0.02; bPop[0] = 0.2; cPop[0] = -65.0; dPop[0] = 8.0; // RS
95     aPop[1] = 0.1; bPop[1] = 0.2; cPop[1] = -65.0; dPop[1] = 2.0; // FS
96     aPop[2] = 0.02; bPop[2] = 0.2; cPop[2] = -50.0; dPop[2] = 2.0; // CH
97     aPop[3] = 0.02; bPop[3] = 0.2; cPop[3] = -55.0; dPop[3] = 4.0; // IB
98     model.initializeSparse();
99
100     float *vPop = model.getScalar<float>("VPop");
101     while(model.getTime() < 200.0f) {
102         model.stepTime();
103         model.pullVarFromDevice("Pop", "V");
104         printf("%f, %f, %f, %f, %f\n", t, vPop[0], vPop[1], vPop[2], vPop[3]);
105     }
106     return EXIT_SUCCESS;
107 }

```

108 2.2 SWIG

109 In order to use GeNN from Python, both the model creation API and the *SharedLibraryModel* 110 functionality need to be 'wrapped' so they can be called from Python. While this is possible using 111 the API built into Python itself, a wrapper function would need to be manually implemented for each 112 GeNN function to be exposed which would result in a lot of maintenance overhead. Instead, we chose 113 to use SWIG (Beazley, 1996) to automatically generate wrapper functions and classes. SWIG generates

114 Python modules based on special interface files which can directly include C++ code as well as special
115 ‘directives’ which control SWIG, for instance:

```
116 %module(package="package") package
117 %include "test.h"
```

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

```
124 %template(addNeuronPopulationLIF) ModelSpec::addNeuronPopulation<NeuronModels::LIF>;
```

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

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

```
135 %feature("director") NeuronModels::Base;
```

136 2.3 PyGeNN

137 While GeNN *could* be used from Python via the wrapper generated using the techniques described in the
138 previous section, the resultant code would be unpleasant to use directly. For example, rather than being
139 able to specify neuron parameters using a native Python data structure such as a list or dictionary, one
140 would have to use a wrapped type such as `DoubleVector([0.25, 10.0, 0.0, 0.0, 20.0, 2.0, 0.5])`. To provide
141 a more user-friendly and pythonic interface, we have built PyGeNN on top of the wrapper generated by
142 SWIG. PyGeNN combines the separate model building and simulation stages of building a GeNN model
143 in C++ into a single API, likely to be more familiar to users of existing Python-based model description
144 languages such as PyNEST (Eppler et al., 2009) or PyNN (Davison et al., 2008). By combining the two
145 stages together, PyGeNN can provide a unified dictionary-based API for initialising homogeneous and
146 heterogeneous parameters as shown in this re-implementation of the previous example:

```
147 from pygenn import genn_wrapper, genn_model
148
149 model = genn_model.GeNNModel("float", "izhikevich")
150 model.dT = 0.1
151
152 izk_init = {"V": -65.0,
153            "U": -20.0,
154            "a": [0.02, 0.1, 0.02, 0.02],
```

```

155         "b": [0.2,      0.2,      0.2,      0.2],
156         "c": [-65.0,    -65.0,    -50.0,    -55.0],
157         "d": [8.0,      2.0,      2.0,      4.0]}
158
159 pop = model.add_neuron_population("Pop", 4, "IzhikevichVariable", {}, izk_init)
160 model.add_current_source("CS", "DC", "Pop", {"amp": 10.0}, {})
161
162 model.build()
163 model.load()
164
165 v = pop.vars["V"].view
166 while model.t < 200.0:
167     model.step_time()
168     model.pull_state_from_device("Pop")
169     print("%t, %f, %f, %f, %f" % (model.t, v[0], v[1], v[2], v[3]))

```

Initialisation of variables with homogeneous values – such as the neurons’ membrane potential – is performed by GeNN and those with heterogeneous values – such as the *a*, *b* and *c* parameters – are initialised by PyGeNN when the model is loaded. While the PyGeNN API is more pythonic and, hopefully, more user-friendly than the C++ interface, it still provides users with the same low-level control over the simulation. Furthermore, by using SWIG’s numpy (Van Der Walt et al., 2011) interface, the host memory allocated by GeNN can be accessed directly from Python using the `pop.vars["V"].view` syntax meaning that no potentially expensive additional copying of data is required. **(TODO: DEFINING NEW NEURON MODELS, PARAMETERS AND VARIABLES)**

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. 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 higher-level 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 data transferred every timestep is typically very small. To address both of these sources of inefficiency, we have added a new data structure to GeNN which stores spike data for many timesteps on device. To reduce the memory required for this data structure and to make its size independent of neural activity, the spikes emitted by a population of *N* neurons in a single simulation timestep are stored in 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 bitfields stored in a circular buffer. Using this approach, even the spiking output of relatively large models, running for many timesteps can 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.

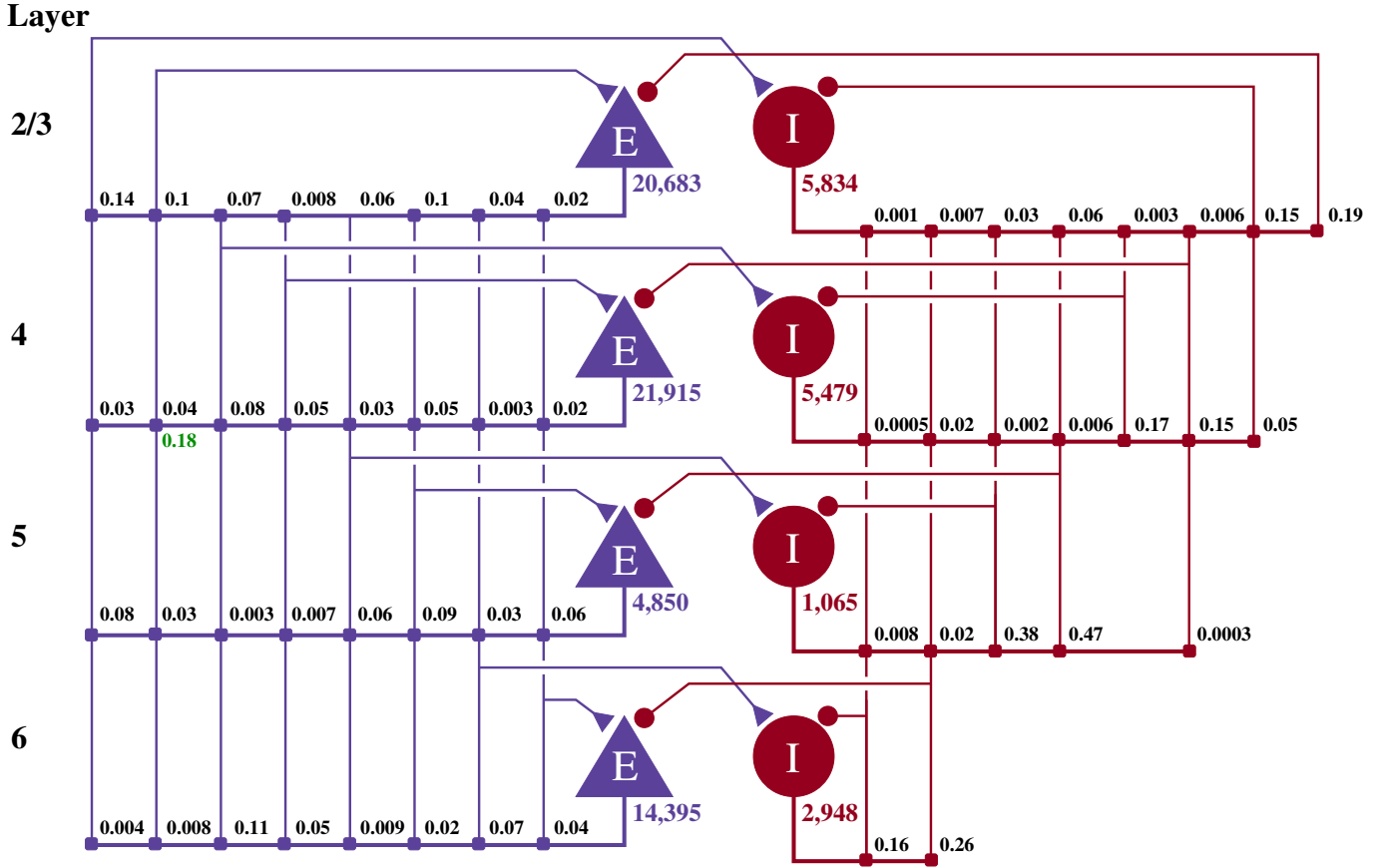


Figure 1. Illustration of the microcircuit model. Blue triangles represent excitatory populations, red circles represent inhibitory populations and the numbers beneath each symbol shows the number of neurons in each population. Connection probabilities are shown in small bold numbers at the appropriate 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 0.008 78 nA. All inhibitory synaptic weights are normally distributed with a mean of 0.3512 nA and a standard deviation of 0.035 12 nA.

199 2.5 Cortical microcircuit model

Potjans and Diesmann (2014) developed a cortical microcircuit model of 1 mm³ of early-sensory cortex. 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) as illustrated by figure 2. 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 $\tau_m = 10$ ms and $R_m = 40$ M Ω represent the time constant and resistance of the neuron's cell membrane, $V_{\text{rest}} = -65$ 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 $V_{\text{th}} = -50$ mV a spike is emitted, the membrane voltage is reset to V_{rest} and updating of V is suspended for a refractory period $\tau_{\text{ref}} = 2$ 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

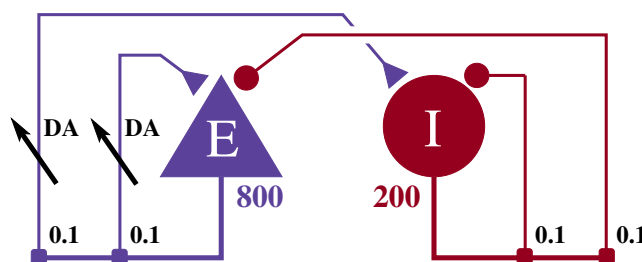


Figure 2. Illustration of the balanced random network model. The blue triangle represents the excitatory population, the red circle represents the inhibitory population, and the numbers beneath each symbol show the number of neurons in each population. Connection probabilities are shown in small bold 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 .

spikes lead to exponentially-decaying input currents I_{syn_i}

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

where $\tau_{\text{syn}} = 0.5$ ms represents the synaptic time constant 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

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

where $\tau_{\text{syn}} = 0.5$ ms, ν_{ext} represents the mean input rate and J represents the weight. The ordinary differential equations 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 timestep of 0.1 ms, making the overheads of copying spikes from the GPU every timestep particularly problematic.

2.6 Pavlovian conditioning model

The cortical microcircuit model described in the previous section is ideal for exploring the performance of short simulations of relatively large models. However, the performance of longer simulations of smaller models is equally vital. (TODO: DETERMINE E.G. PERCENTAGE MODELS E.G. ON OPENSOURCEBRAIN WHICH ARE SMALL). Such models can be particularly troublesome for GPU simulation as, not only might they not offer enough parallelism to fully occupy the device but, each timestep can be simulated so quickly that the overheads of launching kernels etc can dominate. Additional overheads can be incurred when models require injecting external stimuli throughout the simulation. Longer simulations are particularly useful when exploring synaptic plasticity so, to explore the performance

of PyGeNN in this scenario, we simulate a model of Pavlovian conditioning using a three-factor Spike-Timing-Dependent Plasticity (STDP) learning rule (Izhikevich, 2007).

2.6.1 Neuron model

This model 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:

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

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

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

2.6.2 Synapse models

The excitatory and inhibitory neural populations are connected recurrently, as shown in figure 2, 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 based on 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:

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

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

2.6.3 PyGeNN implementation of three-factor STDP

The first step in implementing this learning rule in PyGeNN is to implement the STDP updates and decay of C_{ij} . First, we create a new ‘weight update model’ with the learning rules parameters and the w_{ij} and C_{ij} state variables:

```
234 izhikevich_stdp_model = create_custom_weight_update_class(
235     "izhikevich_stdp",
236
237     param_names=["tauPlus", "tauMinus",
238                 "tauC", "aPlus", "aMinus"],
239     var_name_types=[("w", "scalar"), ("c", "scalar")],
```

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

```
244 is_pre_spike_time_required=True,
245 is_post_spike_time_required=True,
246
247 is_prev_pre_spike_time_required=True,
248 is_prev_post_spike_time_required=True,
```

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

```
252 sim_code=
253     """
254     $(addToInSyn, $(w));
```

Now we need to calculate the time that has elapsed since the last update of C_{ij} using the spike times we previously requested that GeNN record. Within a timestep, GeNN processes presynaptic spikes before

postsynaptic spikes so the time of the last update to C_{ij} will be the latest time either type of spike was processed in previous timesteps:

```
259         const scalar tc = fmax($(prev_sT_pre),
260                                $(prev_sT_post));
```

Using this time, we can now calculate how much to decay C_{ij} following equation 7:

```
262         const scalar tagDecay = exp(-$(t) - tc) / $(tauC));
263         scalar newTag = $(c) * tagDecay;
```

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

```
267         const scalar dt = $(t) - $(sT_post);
268         if (dt > 0) {
269             newTag -= ($(aMinus) * exp(-dt / $(tauMinus)));
270         }
271         $(c) = newTag;
272         """,
```

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

```
277     learn_post_code=
278         """
279         const scalar tc = fmax($(sT_pre),
280                                $(prev_sT_post));
281
282         const scalar tagDecay = exp(-$(t) - tc) / $(tauC));
283         scalar newTag = $(c) * tagDecay;
284
285         const scalar dt = $(t) - $(sT_pre);
286         if (dt > 0) {
287             newTag += ($(aPlus) * exp(-dt / $(tauPlus)));
288         }
289         $(c) = newTag;
290         """)
```

Adding the synaptic weight w_{ij} update described by equation 8 requires two components. In addition to pre and postsynaptic spikes, the weight update model needs to receive events whenever dopamine is injected via DA. **(TODO: IMPROVE SENTANCE)** GeNN supports such events via the ‘spike-like event’ system which allows events to be triggered based on a condition applied to the presynaptic neuron. In this case, this condition is simply used to check an `injectDopamine` flag set by the dopamine injection logic in our presynaptic neuron model:

```
297     event_threshold_condition_code="injectDopamine",
```

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

```
300     is_pre_spike_event_time_required=True,
301     is_prev_pre_spike_event_time_required=True,
```

302 The spike-like events can now be handled using an ‘event code’ string:

```
303     event_code=
304         """
305         const scalar tc = fmax($(sT_pre), fmax($(prev_sT_post), $(prev_seT_pre)));
306         const scalar tagDecay = exp(-$(t) - tc) / $(tauC);
307         $(c) *= tagDecay;
308         """,
```

After updating the previously defined calculations of t_c in the sim code and learn post code to also include the times of spike-like events, all that remains is to update w_{ij} . Mikaitis et al. (2018) showed how equation 8 could be integrated 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 equation 12 can be simplified to:

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

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

311 2.6.4 PyGeNN implementation of Pavlovian conditioning experiment

312 To perform the Pavlovian conditioning experiment using this model, we chose 100 random groups of 50
 313 neurons (each representing stimuli $S_1 \dots S_{100}$) are chosen from amongst the two neural populations. Stimuli
 314 are presented to the network in a random order, separated by intervals sampled from $U(100, 300)$ ms. The
 315 neurons associated with an active stimulus are stimulated for a single 1 ms simulation timestep with a
 316 current of 40.0 nA, in addition to the random background current of $U(-6.5, 6.5)$ nA, delivered to each
 317 neuron via I_{ext_i} throughout the simulation. S_1 is arbitrarily chosen as the Conditional Stimuli (CS) and,
 318 whenever this stimuli is presented, a reward in the form of an increase in dopamine is delivered by setting
 319 $DA(t) = 0.5$ after a delay sampled from $U(0, 1000)$ ms. This delay period is large enough to allow a few
 320 irrelevant stimuli to be presented which act as distractors. The simplest way to implement this stimulation
 321 regime is to add a current source to the excitatory and inhibitory neuron populations which adds the
 322 uniformly-distributed input current to an externally-controllable per-neuron current. In PyGeNN, the
 323 following model can be defined to do just that:

```
324 stim_noise_model = create_custom_current_source_class(
325     "stim_noise",
```

```

326 param_names=["n"],
327 var_name_types=[("iExt", "scalar", VarAccess_READ_ONLY)],
328 injection_code=
329     """
330     $(injectCurrent, $(iExt) + ($(gennrand_uniform) * $(n) * 2.0) - $(n));
331     """

```

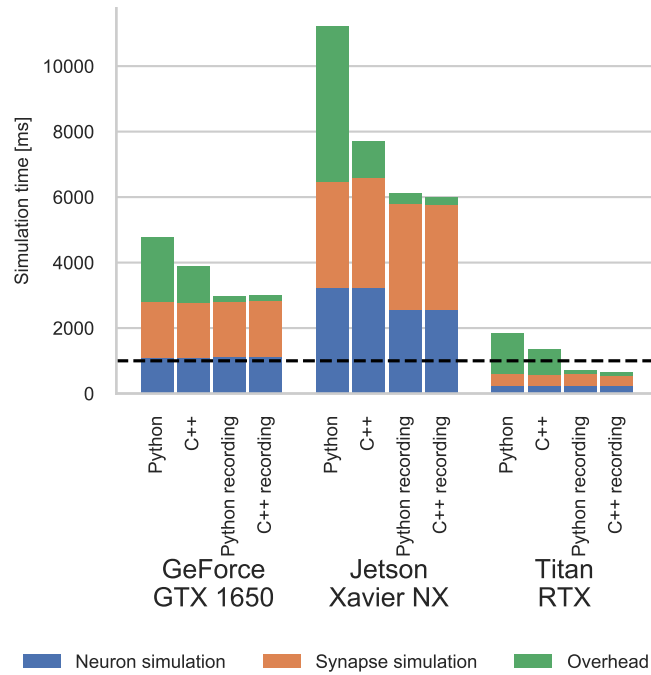


Figure 3. Simulation times of the microcircuit model running on various GPU hardware for 1 s of biological time. ‘Overhead’ refers to time spent in simulation loop but not within CUDA kernels. The dashed horizontal line indicates realtime performance

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

```

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

```

340 The same approach can then be used to zero the current afterwards. However, as almost 20 000 stimuli will
 341 be injected over the course of a 1 h simulation, in order to reduce potential overheads, we can offload the
 342 stimulus delivery entirely to the GPU using the following slightly more complex model:

```

343 stim_noise_model = create_custom_current_source_class(
344     "stim_noise",
345     param_names=["n", "stimMagnitude"],
346     var_name_types=[("startStim", "unsigned int"),

```

```

347         ("endStim", "unsigned int", VarAccess_READ_ONLY)],
348     extra_global_params=[("stimTimes", "scalar*")],
349     injection_code=
350         """
351         scalar current = $(gennrand_uniform) * $(n) * 2.0) - $(n);
352         if($(startStim) != $(endStim) && $(t) >= $(stimTimes)[$(startStim)]) {
353             current += $(stimMagnitude);
354             $(startStim)++;
355         }
356         $(injectCurrent, current);
357         """

```

358 This model retains the same logic for generating background noise but, additionally, uses a simple sparse
359 matrix data structure to store the times at which each neuron should have current injected. **(TODO:**
360 **FIGURE)** The `startStim` and `endStim` variables point to the subset of the `stimTimes` array used by each
361 neuron's current source and, once the simulation time t passes the time pointed to by `startStim`,
362 current is injected and `startStim` is advanced. This array is stored in a 'extra global parameter' which
363 is a read-only memory area that can be allocated and populated from PyGeNN, in this case by 'stacking'
364 together a list of lists of spike times:

```

365 curr_pop.set_extra_global_param("stimTimes", np.hstack(neuron_stimuli_times))

```

3 RESULTS

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

- 368 • Jetson Xavier NX – a low-power embedded system with a GPU based on the Volta architecture with
369 8 GB of shared memory.
- 370 • GeForce GTX 1050Ti – a low-end desktop GPU based on the Pascal architecture with 4 GB of
371 dedicated memory.
- 372 • GeForce GTX 1650 – a low-end desktop GPU based on the Turing architecture with 4 GB of dedicated
373 memory.
- 374 • Titan RTX – a high-end workstation GPU based on the Turing architecture with 24 GB of dedicated
375 memory.

376 All of these systems run Ubuntu 18 apart from the system with the GeForce 1050 Ti which runs Windows
377 10.

378 3.1 Cortical microcircuit model performance

379 Figure 3 shows the simulation times for the full-scale microcircuit model and, as one might predict, the
380 Jetson Xavier NX is slower than the two desktop GPUs. However, considering that it only consumes a
381 maximum of 15 W compared to 75 W or 320 W for the GeForce GTX 1650 and Titan RTX respectively, it
382 still performs impressively.

383 The time taken to actually simulate the models ('Neuron simulation' and 'Synapse simulation') are the
384 same when using Python and C++ as all GeNN optimisation options are exposed to PyGeNN. However,
385 both the PyGeNN and C++ simulations spend a significant amount of every simulation step copying spike

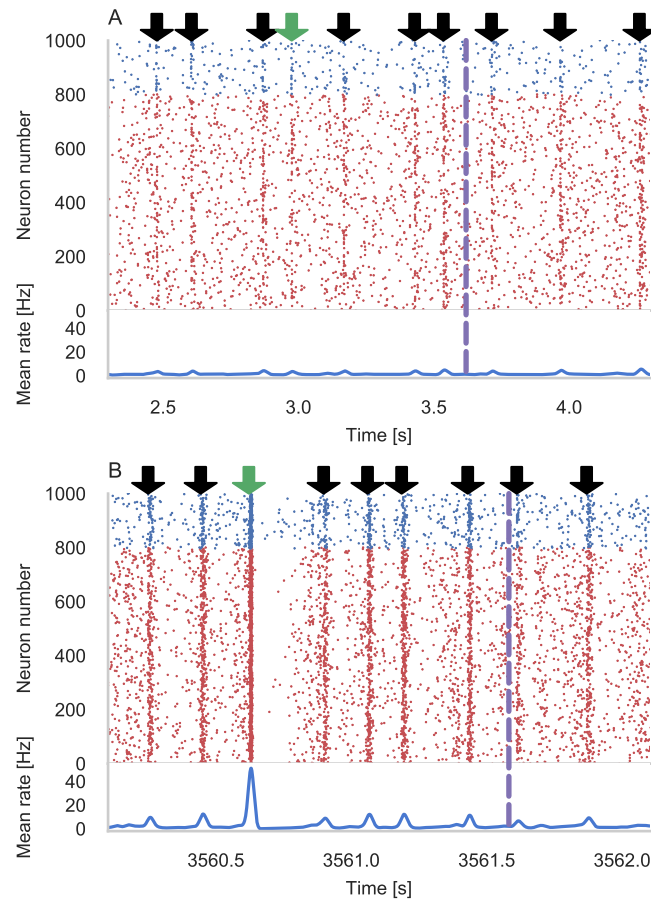


Figure 4. Results of Pavlovian conditioning experiment. Raster and spike density plots showing activity centred around first delivery of Conditional Stimulus (CS) during initial (A) and final (B) 50 s of simulation. Downward green arrows indicate times at which CS is delivered and downward black arrows indicate times when other, un-rewarded stimuli are delivered. Vertical dashed lines indicate times at which dopamine is delivered

data off the device and storing it in a suitable data structure ('Overhead'). Because Python is an interpreted language, such operations are inherently slower – this is particularly noticeable on devices with a slower CPU such as the Jetson Xavier NX. Unlike the desktop GPUs, the Jetson Xavier NX's 8 GB of memory is shared between the GPU and the CPU meaning that data doesn't have to be copied between GPU and CPU memory and can instead be accessed by both. **(TODO: RUN XAVIER NX WITHOUT RECORDING AND WITHOUT ZERO COPY)** While, using this shared memory for recording spikes, reduces the overheads associated with 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. 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 this overhead is reduced by around a factor of 10. Intriguingly, the overhead is now identical between Python and C++. Furthermore, on the high-end desktop GPU, the simulation now runs faster than real-time in both Python and native C++ versions – previously only achievable using a specialised neuromorphic system (Rhodes et al., 2020) and significantly faster than other recently published GPU simulators (Golosio et al., 2020).

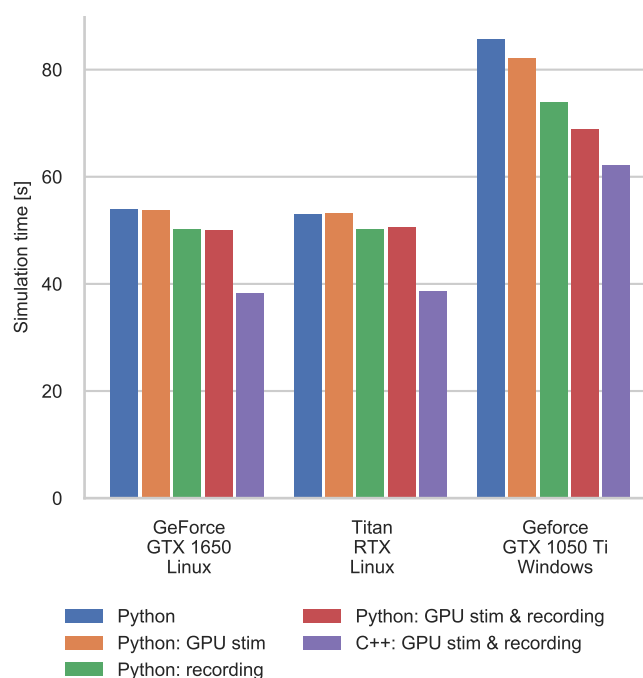


Figure 5. Simulation times of the Pavlovian Conditioning model running on various GPU hardware for 1 h or biological time.

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.

Figure 5 shows the runtime performance for simulations of the Pavlovian conditioning model, running on a selection of desktop GPUs using PyGeNN with and without the recording system described in section 2.4 and the optimized stimuli-delivery described in section 2.6. These PyGeNN results are compared to a C++ simulation which also takes advantage of both optimizations. Interestingly the Titan RTX and GTX 1650 perform identically in this benchmark with speedups ranging from $62\times$ to $72\times$ real-time. This 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. Furthermore, on these two systems, while using the recording system significantly improves performance, the impact of delivering stimuli on the GPU is minimal. However, 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. Furthermore, unlike on the other devices, offloading stimuli delivery to the GPU improves the performance significantly. 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

423 have equal numbers of dedicated integer and floating point ALUs. This is particularly beneficial for
424 SNN simulations which involve a significant amount of integer arithmetic for indexing sparse matrix
425 data structures etc, that is interspersed between the floating point computations needed to determine
426 neuron and synapse states. Furthermore, the large performance improvement seen when offloading
427 stimulus delivery to the GPU is likely to be due to overheads relating to the Windows Display Driver
428 Model (WDDM).**(TODO: CONVINCE NSIGHT SYSTEMS TO WORK AND GET WDDM STATS).**
429 **(TODO: WHY DOES HERE A DIFFERENCE REMAIN BETWEEN PYTHON AND C++ (PRESUMABLY**
430 **IN THE OVERHEADS?; WOULD BE NICE TO BE ABLE TO SEE THE NEURON, SYNAPSE, OVERHEADS**
431 **SPLIT HERE AS WELL!))**

4 DISCUSSION

432 discuss!

- 433 • Turing architecture is great for GeNN! Presented results improve on state-of-the-art.
- 434 • PyGeNN as an intermediate layer - PyNN, ML
- 435 • Cost of C++ - Python calls in models
- 436 • something about neuromorphic systems often being real-time / BS accelerated time

CONFLICT OF INTEREST STATEMENT

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

AUTHOR CONTRIBUTIONS

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

FUNDING

443 This work was funded by the EPSRC (Brains on Board project, grant number EP/P006094/1).

ACKNOWLEDGMENTS

444 This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that
445 aided the efforts of the authors.

DATA AVAILABILITY STATEMENT

446 The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY]
447 [LINK].

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
- 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
- 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
- Davison, A. P., Brüderle, D., Eppler, J., Kremkow, J., Müller, E., Pecevski, D., et al. (2008). PyNN: A Common Interface for Neuronal Network Simulators. *Frontiers in neuroinformatics* 2, 11. doi:10.3389/neuro.11.011.2008
- Eppler, J. M., Helias, M., Müller, E., Diesmann, M., and Gewaltig, M. O. (2009). PyNEST: A convenient interface to the NEST simulator. *Frontiers in Neuroinformatics* 2, 1–12. doi:10.3389/neuro.11.012.2008
- Gewaltig, M.-O. and Diesmann, M. (2007). NEST (NEural Simulation Tool). *Scholarpedia* 2, 1430
- Givon, L. E. and Lazar, A. A. (2016). Neurokernel: An open source platform for emulating the fruit fly brain. *PLOS ONE* 11, 1–25. doi:10.1371/journal.pone.0146581
- Golosio, B., Tiddia, G., De Luca, C., Pastorelli, E., Simula, F., and Paolucci, P. S. (2020). A new GPU library for fast simulation of large-scale networks of spiking neurons , 1–27
- Hines, M. L., Davison, A. P., and Müller, E. (2009). NEURON and Python. *Frontiers in Neuroinformatics* 3, 1–12. doi:10.3389/neuro.11.001.2009
- Hopkins, M. and Furber, S. B. (2015). Accuracy and Efficiency in Fixed-Point Neural ODE Solvers. *Neural computation* 27, 2148–2182
- Humphries, M. D. and Gurney, K. (2007). Solution Methods for a New Class of Simple Model Neurons M. *Neural Computation* 19, 3216–3225. doi:doi:10.1162/neco.2007.19.12.3216
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 90–95. doi:10.1109/MCSE.2007.55
- Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks* 14, 1569–72. doi:10.1109/TNN.2003.820440
- Izhikevich, E. M. (2007). Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling. *Cerebral Cortex* 17, 2443–2452. doi:10.1093/cercor/bhl152
- Knight, J. C. and Nowotny, T. (2018). GPUs Outperform Current HPC and Neuromorphic Solutions in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model. *Frontiers in Neuroscience* 12, 1–19. doi:10.3389/fnins.2018.00941
- Knight, J. C. and Nowotny, T. (2020). Larger GPU-accelerated brain simulations with procedural connectivity. *bioRxiv* doi:10.1101/2020.04.27.063693
- Mikaitis, M., Pineda García, G., Knight, J. C., and Furber, S. B. (2018). Neuromodulated Synaptic Plasticity on the SpiNNaker Neuromorphic System 12, 1–13. doi:10.3389/fnins.2018.00105

- 492 Millman, K. J. and Aivazis, M. (2011). Python for scientists and engineers. *Computing in Science and*
493 *Engineering* 13, 9–12. doi:10.1109/MCSE.2011.36
- 494 Pauli, R., Weidel, P., Kunkel, S., and Morrison, A. (2018). Reproducing Polychronization: A Guide to
495 Maximizing the Reproducibility of Spiking Network Models. *Frontiers in Neuroinformatics* 12, 1–21.
496 doi:10.3389/fninf.2018.00046
- 497 Potjans, T. C. and Diesmann, M. (2014). The Cell-Type Specific Cortical Microcircuit: Relating Structure
498 and Activity in a Full-Scale Spiking Network Model. *Cerebral Cortex* 24, 785–806. doi:10.1093/cercor/
499 bhs358
- 500 Rhodes, O., Peres, L., Rowley, A. G. D., Gait, A., Plana, L. A., Brenninkmeijer, C., et al. (2020). Real-time
501 cortical simulation on neuromorphic hardware. *Philosophical Transactions of the Royal Society A:*
502 *Mathematical, Physical and Engineering Sciences* 378, 20190160. doi:10.1098/rsta.2019.0160
- 503 Stimberg, M., Brette, R., and Goodman, D. F. (2019). Brian 2, an intuitive and efficient neural simulator.
504 *eLife* 8, 1–41. doi:10.7554/eLife.47314
- 505 Stimberg, M., Goodman, D. F., and Nowotny, T. (2020). Brian2GeNN: accelerating spiking neural network
506 simulations with graphics hardware. *Scientific Reports* 10, 1–12. doi:10.1038/s41598-019-54957-7
- 507 Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient
508 numerical computation. *Computing in Science and Engineering* 13, 22–30. doi:10.1109/MCSE.2011.37
- 509 Vitay, J., Dinkelbach, H., and Hamker, F. (2015). ANNarchy: a code generation approach to neural
510 simulations on parallel hardware. *Frontiers in Neuroinformatics* 9, 19. doi:10.3389/fninf.2015.00019
- 511 Yavuz, E., Turner, J., and Nowotny, T. (2016). GeNN: a code generation framework for accelerated brain
512 simulations. *Scientific reports* 6, 18854. doi:10.1038/srep18854