Modules for Automated Validation and Comparison of Models of Neurophysiological and Neurocognitive Biomarkers of Psychiatric Disorders: ASSRUnit - A Case Study

Christoph Metzner, Tuomo Mäki-Marttunen, Achim Schweikard,
Bartosz Zurowski and Volker Steuber

November 23, 2017

Abstract

1 Introduction

Research on psychiatric disorders has gradually shifted its focus from complex clinical phenotypes towards the identification of biomarkers and endophenotypic measures. This change of focus stems from the hope that these measures might map nicely to genetic alterations identified by large genome-wide association studies [30] and promises to more readily shed light on the mechanisms underlying these disorders and to facilitate the discovery of novel medications [49]. Naturally, a lot of effort has been put into the translation of these measures into practice using human studies [40] as well as animal models [27]

Computational approaches also have gained significantly more attention over the last years and this has led to the emergence of 'Computational Psychiatry' as an independent discipline (see e.g. [32, 61, 16, 12, 54, 1]). This emergence can be attributed to three main factors: First, the above mentioned increase in experimental studies has provided a wealth of anatomical and neurophysiological data which are essential to build computational models. Second, methodological and infrastructural advances, such as the various atlases, databases and online tools from the Allen Brain Institute (*http://brain-map.org/*) or the BRAIN initiative (*https://www.braininitiative.nih.gov/*), have made it possible to analyze and process this enormous amount of data. Third, the increase in computing power of high performance computers as well as standard personal computers has made it possible (and affordable) to build and use models of increasingly high computational complexity. Therefore, the rapid growth of the field of computational psychiatry comes as no surprise. However, in order to fully exploit the potential that computational modeling offers, we have to identify systemic weaknesses in current approaches and take a look at other disciplines that use computational models (and have used them for much longer than psychiatry) and even look at disciplines, like software development, which face similar challenges.

At the core of computational modeling lies the concept of validation, i.e. the rigorous comparison of model predictions against experimental findings. Furthermore, for a model to be useful and provide a true contribution to knowledge, the validation has to use sound criteria and the experimental observations need to sufficiently characterize the phenomenon the model tries to reproduce. Hence, in order to develop a computational model scientists need to have an in-depth understanding of the current, relevant experimental data, the current state of computational modeling in the given area and the state-of-the-art of statistical testing, to choose the appropriate criteria with which the model predictions and experimental observations will be compared [17, 46]. In a field where both the number of experimental and computational studies grows rapidly, as is the case for psychiatry, this becomes more and more impracticable. Furthermore, the increase in modeling and experimental studies has made it harder for reviewers not only to judge whether a new model adequately replicates the full range of experimental observations but also how it compares to competing models. Again, also reviewers need an in-depth knowledge of the modeling and experimental literature. Finally, since computational modeling tries to generate predictions which can be experimentally tested, experimentalists must be able to extract and assess predictions from a rapidly growing body of computational models, a task which is also becoming more and more impracticable.

The problems described above are not unique to the field of computational psychiatry but occur in all scientific areas that use computational models. Furthermore, building, i.e. implementing, a computational model is in the end a software development project of sort. Omar et al. [38] have therefore proposed a framework for automated validation of scientific models, SciUnit, which is based on unit testing, a technique commonly used in software development. SciUnit addresses the problems mentioned above by making the scope (i.e. the set of observable quantities that it can generate predictions about) of the model explicit and by allowing its validity (i.e. the extent to which its predictions agree with available experimental observations of those quantities.) to be automatically tested [38].

In this paper, we propose to adopt this framework for the computational psychiatry community and to collaboratively build common repositories of computational models, tests, test suites and tools. As a case in point, we have implemented a Python module (*ASSRUnit*) for auditory steady-state response (ASSR) deficits in schizophrenic patients, which are based on observations from several experimental studies ([22, 57, 23])and we demonstrate how existing computational models ([29, 3, 57, 28]) can be validated against these observations and compared with each other.

2   The SciUnit Framework

The module we present here is based on the general SciUnit framework for the validation of scientific models against experimental observations [38] (see Figure 1).

In SciUnit models declare and implement so-called capabilities, which the validation tests then use to interact with those models. Furthermore, the declaration and implementation of capabilities is separated, which allows to test two different models which share the same capabilities on the same experimental
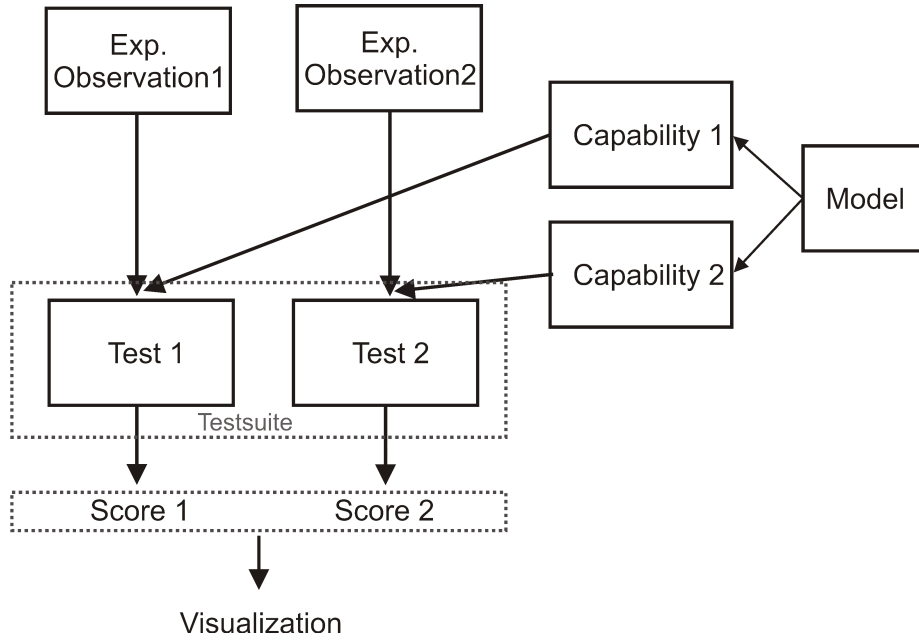
Figure 1: Schematic representation of the SciUnit framework. Models can be tested against experimental observations using specific tests. These tests incorporate an experimental observation and interface with the model through capabilities. Tests can be grouped into so-called test suites. The execution of a test produces a score, which describes how well the model captures the experimental observations. SciUnit also provides methods to visualize the resulting score(s), for example in a table.

observations using the same test. Tests then take the model, use its capabilities to generate data and compare these data to the experimental observations which are linked to the test and create a score. This score, which can simply be a Boolean (pass/fail) or another more complex score type, describes if and to which extent the model data and the experimental observation(s) match.

Before we describe the actual implementations of capabilities, models, tests and scores in our framework for auditory steady-state responses in schizophrenia, we first start with a summary of the experimental observations we included in the database and then we describe the computational models which were realized.

3 The *ASSRUnit* Module

The structure of the module proposed here is shown schematically in Figure 2. As outlined earlier, there are three main functionalities the proposed module aims to provide: 1) To provide a simple way of getting an overview of the experimental literature, 2) To provide an easy and flexible way to automatically test computational models against experimental observations, 3) To provide an automated way of generating predictions from computational models. Functionality 1 is fully covered by the experimental database and its methods to query the database and visualize the results. Functionality 2 is provided by linking both the experimental database as well as the computational models to the SciUnit tests that cover the relevant experimental obervations. The only action required from the user is, if the computational model has not yet been included into the model repository of the module, to provide an interfacing Python class for the model which implements all the required capabilities. Note that the model itself does not have to be written in Python, it only has to be executable from shell. Once the model is included, the SciUnit framework allows for automated testing and the visualization methods provided in the proposed module allow for a comprehensive and clear presentation of the results. Functionality 3 can be achieved by a set of SciUnit tests and capabilites that, instead of covering experimental observations, cover experiments that have not yet been performed. By running the computational models with these tests, the module can be used to generate new predictions from the models, which can then be used to populate a prediction database similar to the experimental database. The module is available on GitHub: *https://github.com/ChristophMetzner/ASSRUnit*

3.1 Experimental Observations Database

In patients suffering from schizophrenia oscillatory deficits in general and ASSR deficits in particular have been extensively studied using EEG and MEG (e.g. [23, 57, 22, 26, 63, 20, 7, 52, 53, 51, 39, 34]). Here, we focus on three of these studies looking at entrainment deficits in the gamma and beta range. Kwon et al. [23] used a click train paradigm to study ASSRs at 20, 30, and 40 Hz in schizophrenic patients using EEG and found a prominent reduction of power at the driving frequency for 40 Hz drive, an increase in power at the driving frequency during 20 Hz drive and no changes for 30 Hz drive. Furthermore, they found small changes of power at certain harmonic/subharmonic frequencies, namely, an increase of power at 20 Hz for 40 Hz drive and a decrease of power
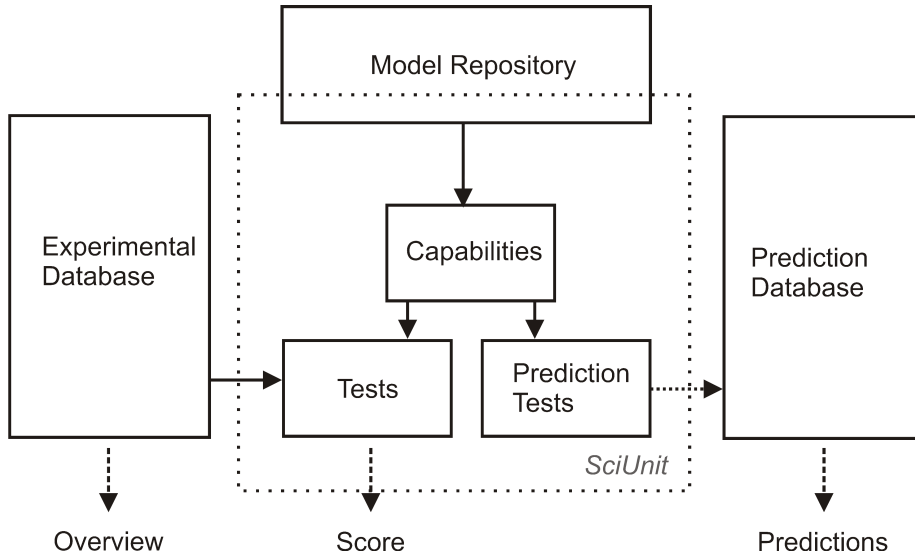
Figure 2: Schematic representation of the proposed framework highlighting the three main functions: 1) Overview of experimental observations. 2) Validation of computational models. 3) Creation of a predictions database. At its core lies the SciUnit module, which provides the infrastructure for the automated validation of the computational models. In particular, through a set of suitable tests, the computational models can be compared against experimental observations queried from the experimental database. Another set of tests, the so-called prediction tests, are then employed to extract predictions from the computational models, thus populating the predictions database.

at 40 Hz for 20 Hz drive. Vierling-Claassen et al. [57] reproduced these findings using the same paradigm with MEG. Krishnan et al. [22] used a slightly different paradigm, which employed amplitude-modulated tones instead of click trains, and tested a wide range of driving frequencies from 5 to 50 Hz. They found reduction of power at the driving frequency in the gamma range (i.e. at 40, 45 and 50 Hz) and no changes at other frequencies. Furthermore, they did not find any changes of power at harmonic or subharmonic frequencies.

The experimental database is realized as a nested Python dictionary, with an entry for each study included. Each study entry consists of two entries, which describe the study observations, one in a quantitative way and the other in a qualitative way. We have included the qualitative description because often either computational models do not allow for a strict quantitative comparison with experimental data or publications of experimental studies do not provide enough detail on the results, and in these cases, only a qualitative comparison is possible.

Table 1: Summary of ASSR deficits in schizophrenic patients in the three studies considered here (↓: sign. lower in patients, ↑: sign. higher in patients, −: no sign. difference between controls and patients). Since Kwon et al. [23] and Vierling-Claassen et al. [57] produced the same results, they are combined here. The tests included in the *ASSRUnit* module are based on this table. Note that Krishnan et al. [22] tested more driving frequencies than the ones shown in the table. The table only shows measures that are common to all three studies.

Drive	Fundamental			Harmonic	Subharmonic
	40 Hz	30 Hz	20 Hz	20 Hz	40 Hz
Kwon/Vierling	↓	-	↑	↓	↑
Krishnan	↓	-	-	-	-

Together with the database, *ASSRUnit* provides basic methods to query and visualize the content of the database. These methods include commands to retrieve all studies or observations in the database and a method to display an overview of the results for the whole database or for certain studies or observations. Finally, the meta-data associated with each study (for example, the number of participants, the modality, the patient group, etc.) can also be retrieved and displayed.

3.2 Prediction Database

The prediction database is also realized as a nested Python dictionary. Similar to the experimental observation database, methods that retrieve and visualize the content of the database are included in *ASSRUnit*.

3.3 Models, Capabilities, Tests and more

Models In order to demonstrate the flexibility of the proposed framework, we included three different neural models of ASSR deficits.

The first model is based on a biophysically detailed model of primary auditory cortex by Beeman [3]. It has recently been used to study ASSR deficits by

our group [29]. The model was implemented using the neural simulator GEN-ESIS [4, 5]. Not only is this model a good example of a biophysically detailed model of ASSR deficits, its inclusion also demonstrates how models that are not written in Python can be used.

The second model is a reimplementation of the model of Beeman in Neu-roML2, a simulator-independent markup language to describe neural network models developed by the NeuroML project [9], which is featured in the open source brain model database [19]. We included this model to demonstrate the ability of the proposed framework to incorporate state-of-the-art tools and databases for the design, implementation and simulation of network models.

The last model we included is the simple model presented by Vierling-Claassen et al. [57]. The model is a simple network of two populations of theta neurons. We reimplemented the model in Python (for more details on the model and the replication see [28]). The model was included first of all to demonstrate that the framework is not limited to biophysically detailed models but can also be used with simpler, more abstract models. Additionally, the inclusion of the model demonstrates the simplest way of including a model, implementing the model in Python. This might not be the most common scenario, but since it is the simplest, we included it here.

We do not discuss the models in more detail here, since they have been described elsewhere [3, 29, 57, 28]. Furthermore, our focus lies on the framework with which to use, validate and compare models not on the models themselves.

The three models mentioned above are included into the SciUnit framework by wrapper classes that implement the necessary capabilities and make the models available to the tests. One important thing to note here, is that, since we are dealing with models of deficits in disorders, that a 'model' as used in the module, always means two configurations of a computational model, one representing the control configuration and one the disorder configuration. Therefore, all wrapper classes take two sets of parameters as an argument describing the necessary parameters for the two configurations, respectively. In addition to the standard model classes, we also implemented a second version of the model classes, which simulates a certain number n of simulations, instead of a single one, where each simulation differs in background noise. This allows for assessing the robustness of the results.

Capabilities   Table 3.1 summarizes the experimental observations included in the module at this stage. All observations are similar in nature: the power value of the EEG/MEG at a certain frequency in response to auditory entrainment at a certain frequency. Therefore, the only capability necessary for a model to produce output that can be compared to these observations is a method that produces the power at a certain frequency X of a simulated EEG/MEG signal in response to drive at a frequency Y. This capability, *ProduceXY*, is inlcuded in ASSRUnit and all models must implement it.

Tests and Scores   Since tests The five tests we implemented, test the five observations summarized in Table 3.1 individually. Furthermore, we implemented one prediction test, which tests 10 Hz power at 10 Hz drive. For the sake of simplicity, the test scores implemented so far are simple Boolean score, indicating whether a model output fails or passes a test, i.e. whether the difference between

List all studies in database

```
In [2]: s=get_studies()
        print s

['Kwon_1999', 'Krishnan_2009', 'Vierling_2008']
```

List all observations in database

```
In [3]: o=get_observations()
        print o

['2525 :25 Hz power at 25 Hz drive', '2040 :20 Hz power at 40 Hz drive',
 '1010 :10 Hz power at 10 Hz drive', '3030 :30 Hz power at 30 Hz drive',
 '2020 :20 Hz power at 20 Hz drive', '1515 :15 Hz power at 15 Hz drive',
 '4545 :45 Hz power at 45 Hz drive', '4040 :40 Hz power at 40 Hz drive',
 '0505 :5 Hz power at 5 Hz drive', '4020 :40 Hz power at 20 Hz drive',
 '3535 :35 Hz power at 35 Hz drive', '5050 :50 Hz power at 50 Hz drive']
```

Figure 3: Display all studies and all observations included in the database

model output for the control and the 'schizophrenia-like' network matches the experimental observation. In case of the model classes implementing sets of outputs, simply the mean difference is compared to the experimental observations. The only test, where the score is not a Boolean, is the prediction test, since here we do not have an experimental value to compare to. Here, we have chosen a RatioScore, which returns the ratio of the power for the 'schizophrenia-like' configuration and the power for the control configuration.

Visualization, Statistics, Additional Data In addition to the main features of the SciUnit framework for the analysis and comparison of the models, we use the fact that SciUnit allows to pass additional data, beyond the test scores, to provide a class that offers tools for the visualization of the results. This class includes functions to display the test results in a table, to plot the test results as a bar graph, to plot the results from a set of model outputs as a box plot, to perform and visualize a student's t-test of the differences between control and 'schizophrenia-like' networks

Next, we describe three different use cases, which show how the proposed module can be used for different purposes by experimentalists, modelers and reviewers.

3.4 Use Case I: Overview of the Experimental Literature

The first use case demonstrates how the experimental database can be used to get a comprehensive overview of the current experimental literature related to a neurophysiological or neurocognitive biomarker, in our case auditory steady-state response deficits in patients suffering from schizophrenia. Figure 3 shows that with two simple commands one can retrieve the names of all studies and all observations present in the database. These names will have to be used for all further queries of the database.

Figure 4 a) then shows how to get a complete overview of all observations of all studies in the database. As we can see in Figure 4 b), simply adding

the parameter *meta=true*, to the command, will additionally output the meta-data associated with each study. This contains information on the subjects, modality and so on. The overview command presents the data in a simple table and can be used to see which studies provided which observation and what the results were. However, as we can already see for our small demonstration database containing only three studies, this is likely to become big and therefore hard to fully grasp. By explicitly stating the studies and/or the observations one is interested in, one can reduce the complexity of the table and get a clear and simple overview, as depicted in Figure 5. Note that in the examples, we have only used the qualitative description of the observations, the same functionality also applies to the quantitative descriptions. The functionality described here, along with more examples, can be explored in an accompanying Python notebook (*Example_Experimental_Database.ipynb* in *https://github.com/ChristophMetzner/ASSRUnit/Code/*).

This simple querying functionality allows the user to get a quick, clean and comprehensive overview of the experimental literature, to identify observations that are supported by many studies (see in our case the reduction of gamma power for stimulation at gamma frequency) but also to detect controversial findings. Furthermore, the display of the associated meta-data allows to check for example whether identified common observations extend over different modalities and post-processing techniques, and also whether controversial findings might be explained by differences in the experimental setup or other related aspects. In the future, it will also be possible to look at more than one database and compare the same observations across different patient groups to highlight commonalities and differences between disorders.

3.5   Use Case II: Model Comparisons

While our first use case only exploited the experimental database, we now show the additional benefits of joining experimental and modeling data.

Simple model comparison   By creating tests, based on the models capabilities, and grouping them into test suites, we can easily compare models against experimental data and against each other. Figure 6 demonstrates how we can use the module to create two different models along with several tests, then run the models to produce the data relevant for the tests and then judge the model outputs against experimental data and display the result together. Note that in this context we use model as the *in silico* instantiation of a theoretical/conceptual model. Two different models therefore, do not necessarily have to use different model implementations but might simply differ in parameters.

Advanced modeling data and visualization   As already described in the Methods section, the model classes do not only contain the standard methods that implement the necessary capabilities, but also contain so-called '..._plus' methods which generate additional model data. Together with the methods from the visualization class, this additional model data can be used to better understand the model behavior, to judge the robustness of findings and to statistically analyze model output. Figure 7

```
In [4]: experimental_overview()
```

	Kwon_1999	Krishnan_2009	Vierling_2008
10 Hz power at 10 Hz drive	not tested	equal	not tested
45 Hz power at 45 Hz drive	lower	lower	lower
20 Hz power at 40 Hz drive	higher	equal	higher
35 Hz power at 35 Hz drive	not tested	equal	not tested
40 Hz power at 40 Hz drive	equal	equal	equal
25 Hz power at 25 Hz drive	not tested	lower	not tested
20 Hz power at 20 Hz drive	higher	equal	higher
30 Hz power at 30 Hz drive	lower	equal	lower
15 Hz power at 15 Hz drive	not tested	equal	not tested
40 Hz power at 20 Hz drive	not tested	equal	not tested
50 Hz power at 50 Hz drive	not tested	lower	not tested
5 Hz power at 5 Hz drive	not tested	equal	not tested

(a) Full overview

```
In [8]: a=experimental_overview(meta=True)

Kwon_1999
{   'Comments': 'Values estimated from figures, since values are not provided',
    'Measure': {   'Location': 'Midline frontal electrode',
                   'Modality': 'EEG',
                   'Processing': 'Butterworth bandpass-filtered time averages
                                  followed by Fourier transform',
                   'Value': 'Mean Absolute Power'},
    'Number of subjects': '15 (Ctrl) and 15 (Scz)',
    'Paradigm': 'Click-train',
    'Subjects': 'Schizophrenia vs Control'}
Krishnan_2009
{   'Measure': {   'Location': 'Cz in a 10-20 setting',
                   'Modality': 'EEG',
                   'Processing': 'Least square linear FIR filtered and
                                  Hilbert transformed',
                   'Value': 'Mean baseline corrected power'},
    'Number of subjects': '21 (Ctrl) vs 21 (Scz)',
    'Paradigm': 'Amplitude-modulated tones; carrier freqeuncy 1kHz',
    'Subjects': 'Schizophrenia vs Control'}
Vierling_2008
{   'Comments': 'Values estimated from figures, since values are not provided',
    'Measure': {   'Location': 'Left hemisphere',
                   'Modality': 'MEG',
                   'Processing': 'Time-averaging followed by PSD using
                                  Welch`s method',
                   'Value': 'Mean Absolute Power'},
    'Number of Subjects': '12 (Ctrl) and 12 (Scz)',
    'Paradigm': 'Click-train',
    'Subjects': 'Schizophrenia vs Control'}
```

(b) Including meta data

Figure 4: Overviews of the observations in the experimental literature. a) The command *experimental_overview* prints a table summarizing the results for all studies and all observations in the database. b) By setting the *meta* flag to *True*, additional information on the studies are displayed. Note that by default the qualitative study results are presented. This can be changed to the quantitative results setting the parameter *entrytype* to *Full*.

1.1.4 Specific observations for specific studies

```
In [7]: observations = ['2020','3030','4040']
        studies = ['Kwon_1999','Krishnan_2009']
        experimental_overview(studies=studies,observations=observations)
```

	Kwon_1999	Krishnan_2009
20 Hz power at 20 Hz drive	higher	equal
30 Hz power at 30 Hz drive	equal	equal
40 Hz power at 40 Hz drive	lower	lower

Figure 5: The *experimental_overview* command allows for querying for specific studies and observations using the names retrieved with the *get_studies* and *get_observations* commands.

Model instances

```
In [3]: conceptual_model_1 = VierlingSimpleModel(controlparams_model_1,
                          schizparams_model_1,name='Conceptual_model_1')
        conceptual_model_2 = VierlingSimpleModel(controlparams_model_2,
                          schizparams_model_2,name='Conceptual_model_2')
```

(a) Create model instances

Tests

```
In [4]: test_4040 = Test4040(observation={'ratio':0.5})
        test_3030 = Test3030(observation={'ratio':1.0})
        test_2020 = Test2020(observation={'ratio':1.0})
        test_2040 = Test2040(observation={'ratio':1.0})
        test_4020 = Test4020(observation={'ratio':1.0})
```

(b) Create tests

A test suite

```
In [5]: kwon_vierling_main_suite = sciunit.TestSuite('kwon_vierling_main',
          [test_4040,test_3030,test_2020,test_4020,test_2040])
        score_matrix = kwon_vierling_main_suite.judge([conceptual_model_1,
          conceptual_model_2])
        score_matrix.view()
```

(c) Create a testsuite and run models against it

	Reduction of 40Hz power to 40Hz drive	No change of 30Hz power to 30Hz drive	Increase of 20Hz power to 20Hz drive	Decrease of 40Hz power to 20Hz drive	Increase of 20Hz power to 40Hz drive
Conceptual_model_1	Pass	Fail	Pass	Fail	Pass
Conceptual_model_2	Fail	Pass	Pass	Fail	Fail

(d) Display comparison

Figure 6: Two models compared against each other and experimental observations. a) Model instances are created. b) Appropriate tests are created. c) Tests are grouped together to form a test suite, then the models are run against the test suite. d) A comparison table shows the performance of each model against each test.

11

```
In [3]: test_model = VierlingSimpleModelRobust(controlparams,schizparams,seeds)
```

0.1.3   Run simulations

```
In [4]: print 'Run simulations (this might take 15-20 minutes)'
        print '\n 4040'
        mcontrol4040,mschiz4040,control4040,schiz4040 =
        test_model.produce_4040_plus()
        print '\n 3030'
        mcontrol3030,mschiz3030,control3030,schiz3030 =
        test_model.produce_3030_plus()
        print '\n 2020'
        mcontrol2020,mschiz2020,control2020,schiz2020 =
        test_model.produce_2020_plus()
        print '\n 2040'
        mcontrol2040,mschiz2040,control2040,schiz2040 =
        test_model.produce_2040_plus()
        print '\n 4020'
        mcontrol4020,mschiz4020,control4020,schiz4020 =
        test_model.produce_4020_plus()
```

(a) Create model instances and run simulation
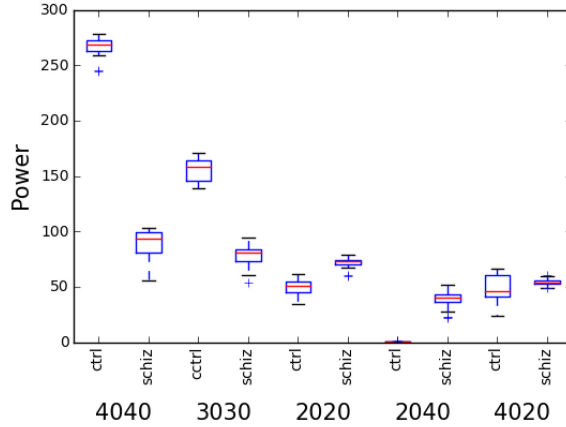
(b) Boxplot of model data

```
In [11]: t4040,p4040 = ttest_ind(control4040,schiz4040)
         t3030,p3030 = ttest_ind(control3030,schiz3030)
         t2020,p2020 = ttest_ind(control2020,schiz2020)
         t2040,p2040 = ttest_ind(control2040,schiz2040)
         t4020,p4020 = ttest_ind(control4020,schiz4020)
         print '40Hz power at 40Hz drive: F=',t4040,'p=',p4040
         print '30Hz power at 30Hz drive: F=',t3030,'p=',p3030
         print '20Hz power at 20Hz drive: F=',t2020,'p=',p2020
         print '20Hz power at 40Hz drive: F=',t2040,'p=',p2040
         print '40Hz power at 20Hz drive: F=',t4020,'p=',p4020

40Hz power at 40Hz drive: F= 51.8698283512 p= 7.00615973185e-37
30Hz power at 30Hz drive: F= 23.4080401139 p= 3.545942616e-24
20Hz power at 20Hz drive: F= -11.3292852332 p= 9.5206066581e-14
20Hz power at 40Hz drive: F= -19.198563914 p= 3.72735850992e-21
40Hz power at 20Hz drive: F= -2.08532975168 p= 0.0438080382856
```

(c) Statistical analysis of model data

Figure 7: Generating additional data with the '..._plus' methods of the model classes. a) Model instances are created and the '..._plus' method is used to run the simulation. b) The additional data visualized as a boxplot. c) Statistical analysis of the additional model data.

Instantiate model

```
In [3]: conceptual_model = VierlingSimpleModel(controlparams_model,
                           schizparams_model,name='Conceptual_model')
```

Create test

```
In [4]: prediction_test_1010 = PredictionTest1010()
```

Use 'judge' method to simulate the model and create the prediction

```
In [5]: score = prediction_test_1010.judge(conceptual_model)
```

(a) Create model instances and tests, and run model

```
In [6]: print score

Ratio = 1.85
```

(b) Display results

Figure 8: An overview of predictions from a model.

3.6 Use Case III: Overview over Model Predictions

Finally, we show how predictions can be generated from existing models (see Figure 7). In order to generate the predictions, a set of prediction tests along with prediction capabilities, that is, capabilities the models must have in order for the model to generate the relevant data needs to be created. For demonstration purposes, we have chosen to implement a single, simple prediction test. Since in *ASSRUnit* so far, we have only looked at experimental observations and computational models that cover gamma and beta range entrainment, the first test simply generates a prediction how, in a given model, power in the alpha band (here at 10 Hz) differs between the control network and the schizrenia-like network at 10 Hz drive. Note that this prediction test has been studied in the experimental literature, which means that it could have already been included in the experimental database and therefore does not represent a true prediction. However, we have chosen to include it for the purpose of demonstration.

4 Discussion

The potential role of the framework within computational psychiatry   The use of computational approaches has seen a significant increase over the last decades in almost all areas of medicine and life sciences. Especially in psychiatry it has become clear that the complex and often polygenic nature of psychiatric disorders can only be understood with the help of computational models [1, 61, 16, 12, 54, 32, 49]. Naturally, the number of computational models in the field of psychiatry has also increased significantly over the last years and it has been argued that *in silico* instantiations of biomarkers and endophenotypes are a crucial step towards an understanding of underlying disease mechanisms [49]. While, this large increase in modeling studies shows the importance of computational methods in the field, it also raises several issues that, impede the community to exploit these approaches to their full potential. In order for a computational model to be a substantial contribution to knowledge it has to ad-

equately instantiate experimental observations, correctly implement the mathematical equations of the model and generate experimentally testable predictions. The approach presented here, addresses two of these three requirements, namely, the instantiation of experimental observations and the generation of testable predictions. While correctness of the code is an equally important requirement, it was out of scope of the current work, since it very strongly depends on the type of computational model and on the programming language used to implement the model. Nevertheless, the approach presented here offers significant benefits for, not only the computational psychiatry community, but for the psychiatry community as a whole, while imposing little additional efforts for the users and contributors. It gives modelers a tool to query experimental observations on neurophysiological and neurocognitive biomarkers, and therefore, helps them to include the current relevant experimental data into their modeling efforts. It further enables them to validate their modeling output against experimental observations during model construction and to demonstrate the performance of their model, both, with respect to the experimental literature and with respect to other competing models. In addition to the benefits it offers the modelers it also enables experimentalists to quickly gain insight into the current state of modeling and to extract experimentally testable predictions from the models. Last but not least, it offers a tool to reviewers which allows them to judge a newly proposed model by making explicit its performance against experimental data and competing models.

The concept of automated code testing and validation has been successfully applied in computer science for many years now, however, it is only slowly finding its way into the computational branches of scientific fields. SciUnit attempts to satisfy this demand by providing a simple, flexible yet powerful framework to address the above-mentioned issues. The computational neuroscience community has started to adopt this framework for the automatic validation of single neuron models (NeuronUnit, [18]). To the best of our knowledge, we are not aware of any similar efforts in the field of psychiatry.

Since schizophrenia is a polygenic, multi-factorial and very heterogeneous disorder, it has been argued that the usefulness of biomarkers and endophenotypes lies in their potential to dissect the disorder into subtypes, which might even be linked more closely to findings on the genetic level [30, 40, 27]. The proposed *ASSRUnit* module together with computational models of biomarkers/endophenotypes and specifically designed test suites could strongly facilitate this process by providing mechanistic links between neurophysiological or neurocognitive biomarkers and changes at the synaptic, cellular and/or network level.

Future directions for ASSRUnit The presented *ASSRUnit* module can be easily extended and modified by others to fit their needs (for example to include more and more specialized visualization tools). Our efforts for establishing *ASSRUnit* as a widely used tool will focus on three main areas: 1) We aim to cover the majority of existing experimental studies with our experimental database in the future. Furthermore, we hope to convince experimentalists to provide more detailed experimental data or to even create database entries themselves. 2) We also aim to cover the majority of current computational models. Again, we hope to encourage modelers to actively contribute to *ASSRUnit*. 3) We aim

to extend our set of prediction tests, and thus, our prediction database.

The most straightforward extension, in our view, is to include information on phase-locking in addition to pure power in certain frequency bands. Several studies report, additionally to a reduction in gamma power, a reduction in the phase-locking factor for patients suffering from schizophrenia (for example [23, 7, 26, 57, 22]). These observations can very easily be incorporated into the existing module, simply by including the experimental observations into the database, adding the necessary capabilities to the model classes and by adding the appropriate tests that link the experimental observations to the model capabilities.

Furthermore, the changes in oscillatory activity upon auditory stimulation are not limited to the gamma and the beta range for schizophrenic patients, but also extend to lower frequency bands such as alpha, theta and delta. For example, Brockhaus-Dumke and colleagues find reduced phase-locking in the alpha and theta band for schizophrenic patients in an auditory paired-click paradigm [8], and Ford et al. find a reduction of phase-locking in the delta and theta range for schizophrenic patients in an auditory oddball task [15]. Abnormalities in these frequency bands have also been found in many other paradigms outside of the auditory system (see [2]). To the best of our knowledge, ASSRs to entrainment stimuli in the theta and delta range have not been looked at in schizophrenia. Therefore, *ASSRUnit* could be used to generate predictions in these frequency ranges as demonstrated in use case III.

However, an inclusion of the above-mentioned observations together with computational models explaining these deficits is not straightforward, because either the paradigms are different from the ones used to elicit ASSRs and/or the mechanisms underlying the effect, and therefore the computational models, are substantially different to ASSRs. Therefore, these deficits are better explored in separate modules solely focusing on each paradigm/deficit. However, it would be very interesting to 'co-explore' computational models that have the capabilities to explain both, ASSR gamma/beta band and delta/theta/alpha phase-locking, deficits. Such an analysis could highlight interactions between different mechanisms underlying different symptoms/biomarkers.

Another very interesting and promising extension of the current module would be to include data and models from different psychiatric disorders, since schizophrenia is not the only disorder where patients show entrainment deficits. Wilson et al. [62], explored gamma power adolescents with psychosis and found reductions compared to normally developing controls. Their patient group consisted of patients suffering from schizophrenia and also from schizoaffective disorder and bipolar disorder. Other studies have also found reduced power and phase-locking in the gamma range in patients with bipolar disorder [37, 53, 43]. The presented module is perfectly suited to highlight commonalities and differences across disorders and to link those to mechanistic explanations via different theoretical/computational models.

Other modules beyond ASSRUnit   The approach presented here, combining an experimental database with a collection of models, tests, prediction tests and a resulting predictions database, can be readily applied to a number of other neurophysiological biomarkers of schizophrenia as well as other psychiatric disorders. In patients suffering from schizophrenia a dysfunction of the auditory

system has long been suspected. In fact, a large number of biomarkers and endophenotypes for schizophrenia, other than ASSR deficits, involve auditory processing. Several alterations of event-related potentials (ERPs) such as mismatch negativity (MMN), N100, and P50 have been described in the literature (see also [49, 48]).

MMN is a negative component of the auditory evoked potential, which is evoked by an alteration in a repetitive sequence of auditory stimuli. MMN seems to be specific to schizophrenia because patients suffering from other psychiatric disorders (for example bipolar disorder and major depression) show normal MMN [55]. Auditory MMN is likely to be generated in primary and secondary auditory cortices and is therefore very similar to stimulus-specific adaptation (SSA) properties of single neurons in auditory cortex [36] (although not identical; see efor example [14, 58]). Several models explaining mechanisms underlying MMN/SSA have been proposed (for example [31, 35]).

The P50 potential, a small positive deflection of the EEG signal at around 50 ms after the onset of an auditory stimulus, is often reduced to the second of a pair of stimuli. However, this so-called P50 reduction is markedly reduced in schizophrenic patients (for example [6]). Another important measure of sensory gating is the pre-pulse inhibition (PPI) of the auditory startle reflex (i.e. the phenomenon in which a weaker prestimulus inhibits the reaction to a subsequent strong startling stimulus). As P50 reduction, PPI is also reduced in schizophrenic patients, although both do not seem to be convergent [6]. Again, computational models exploring the mechanisms underlying PPI have been developed (for example [47, 42, 25], and Moxon et al. have investigated the dopaminergic modulation of the P50 auditory-evoked potential and its relationship to sensory gating in schizophrenia [33].

The N100 (also called N1) is a negative component of the EEG signal occurring approximately 100 ms after stimulus onset. This negative deflection is again reduced in schizophrenic patients (reviewed in [45, 44, 21]) and these deficits have also been modeled [56].

Outside of the auditory system, working memory deficits are probably one of the most robust and best described cognitive deficits in schizophrenic patients (reviewed in [41, 24]). Patients show a decrease in working memory capacity, i.e. the capacity to retain information online for a relatively short period of time, across a broad range of paradigms. Again, several theoretical and computational models have been proposed, aiming to provide mechanistic descriptions of the underlying mechanisms (for example [11, 13, 59, 50, 60, 10]).

All these deficits and alterations along with the mentioned computational models could be integrated into a module similar to the proposed ASSRUnit module.

5   Conclusion

We have proposed a framework for automated validation and comparison of computational models of neurophysiological and neurocognitive biomarkers of psychiatric disorders. The approach builds on SciUnit, a Python framework for scientific model comparison. As case in point, we used this framework to develop *ASSRUnit*, a module comprising an experimental observations data base, computational models, capabilities, tests/test suites and visualization functions

for ASSR response deficits in schizophrenia.

Our approach will facilitate the development, validation and comparison of computational models of neurophysiological and neurocognitive biomarkers of psychiatric disorders by making the scope of models explicit and by making it easy for the user to assess a model's validity and to compare a model against competing models. Furthermore, it is easy to use, straightforward to extend to more experimental observations, computational models and analyses and, ready to apply to other biomarkers. Therefore, the adoption of the proposed framework could be of great use for modelers, reviewers and experimentalists in the field of computational psychiatry.

References

[1] Rick A Adams, Quentin JM Huys, and Jonathan P Roiser. Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(1):53–63, 2016.

[2] Erol Basar and Bahar Guntekin. Review of delta, theta, alpha, beta, and gamma response oscillations in neuropsychiatric disorders. *Suppl Clin Neurophysiol*, 62:303–341, 2013.

[3] David Beeman. A modeling study of cortical waves in primary auditory cortex. *BMC Neuroscience*, 14(Suppl 1):P23, 2013.

[4] James M Bower. Modeling the nervous system. *Trends in Neuroscience*, 1992.

[5] James M Bower and David Beeman. *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* Springer, 1993.

[6] David L Braff, Robert Freedman, Nicholas J Schork, and Irving I Gottesman. Deconstructing schizophrenia: an overview of the use of endophenotypes in order to understand a complex disorder. *Schizophrenia bulletin*, 33(1):21, 2007.

[7] Colleen A Brenner, Olaf Sporns, Paul H Lysaker, and Brian F O'Donnell. Eeg synchronization to modulated auditory tones in schizophrenia, schizoaffective disorder, and schizotypal personality disorder. *American Journal of Psychiatry*, 160(12):2238–2240, 2003.

[8] Anke Brockhaus-Dumke, Ralf Mueller, Ulrich Faigle, and Joachim Klosterkoetter. Sensory gating revisited: relation between brain oscillations and auditory evoked potentials in schizophrenia. *Schizophrenia research*, 99(1):238–249, 2008.

[9] Robert C Cannon, Padraig Gleeson, Sharon Crook, Gautham Ganapathy, Boris Marin, Eugenio Piasini, and R Angus Silver. Lems: a language for expressing complex biological models in concise and hierarchical form and its use in underpinning neuroml 2. *Frontiers in neuroinformatics*, 8, 2014.

[10] Maria Cano-Colino and Albert Compte. A computational model for spatial working memory deficits in schizophrenia. *Pharmacopsychiatry*, 45(S 01):S49–S56, 2012.

[11] Albert Compte, Nicolas Brunel, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923, 2000.

[12] Philip R Corlett and Paul C Fletcher. Computational psychiatry: a rosetta stone linking the brain to mental illness. *The Lancet Psychiatry*, 1(5):399–402, 2014.

[13] Daniel Durstewitz, Jeremy K Seamans, and Terrence J Sejnowski. Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of neurophysiology*, 83(3):1733–1750, 2000.

[14] Brandon J Farley, Michael C Quirk, James J Doherty, and Edward P Christian. Stimulus-specific adaptation in auditory cortex is an nmda-independent process distinct from the sensory novelty encoded by the mismatch negativity. *Journal of Neuroscience*, 30(49):16475–16484, 2010.

[15] Judith M Ford, Brian J Roach, Ralph S Hoffman, and Daniel H Mathalon. The dependence of p300 amplitude on gamma synchrony breaks down in schizophrenia. *Brain research*, 1235:133–142, 2008.

[16] Karl J Friston, Klaas Enno Stephan, Read Montague, and Raymond J Dolan. Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, 1(2):148–158, 2014.

[17] RC Gerkin and C Omar. Neurounit: Validation tests for neuroscience models. In *Front. Neuroinform. Conference Abstract: Neuroinformatics*, 2013.

[18] Richard C Gerkin and Cyrus Omar. Collaboratively testing the validity of neuroscientific models. *Frontiers in Neuroinformatics*, page 1, 2014.

[19] Padraig Gleeson, Eugenio Piasini, Sharon Crook, Robert Cannon, Volker Steuber, Dieter Jaeger, Sergio Solinas, Egidio D'Angelo, and R Angus Silver. The open source brain initiative: enabling collaborative modelling in computational neuroscience. *BMC neuroscience*, 13(1):O7, 2012.

[20] Jordan P Hamm, Anastasia M Bobilev, Lauren K Hayrynen, Matthew E Hudgens-Haney, William T Oliver, David A Parker, Jennifer E McDowell, Peter A Buckley, and Brett A Clementz. Stimulus train duration but not attention moderates γ-band entrainment abnormalities in schizophrenia. *Schizophrenia research*, 165(1):97–102, 2015.

[21] Daniel C Javitt, Kevin M Spencer, Gunvant K Thaker, Georg Winterer, and Mihály Hajós. Neurophysiological biomarkers for drug development in schizophrenia. *Nature reviews. Drug discovery*, 7(1):68, 2008.

[22] GP Krishnan, William P Hetrick, CA Brenner, A Shekhar, AN Steffen, and Brian F O'Donnell. Steady state and induced auditory gamma deficits in schizophrenia. *Neuroimage*, 47(4):1711–1719, 2009.

[23] Jun Soo Kwon, Brian F O'Donnell, Gene V Wallenstein, Robert W Greene, Yoshio Hirayasu, Paul G Nestor, Michael E Hasselmo, Geoffrey F Potts, Martha E Shenton, and Robert W McCarley. Gamma frequency–range abnormalities to auditory stimulation in schizophrenia. *Archives of General Psychiatry*, 56(11):1001–1005, 1999.

[24] Junghee Lee and Sohee Park. Working memory impairments in schizophrenia: a meta-analysis. *Journal of abnormal psychology*, 114(4):599, 2005.

[25] Lorenz Leumann, Désirée Sterchi, Franz Vollenweider, Katja Ludewig, and Hansruedi Früh. A neural network approach to the acoustic startle reflex and prepulse inhibition. *Brain research bulletin*, 56(2):101–110, 2001.

[26] Gregory A Light, Jung Lung Hsu, Ming H Hsieh, Katrin Meyer-Gomes, Joyce Sprock, Neal R Swerdlow, and David L Braff. Gamma band oscillations reveal neural network cortical coherence dysfunction in schizophrenia patients. *Biological psychiatry*, 60(11):1231–1240, 2006.

[27] Athina Markou, Christian Chiamulera, Mark A Geyer, Mark Tricklebank, and Thomas Steckler. Removing obstacles in neuroscience drug discovery: the future path for animal models. *Neuropsychopharmacology*, 34(1):74–89, 2009.

[28] Christoph Metzner. [re] modeling gaba alterations in schizophrenia: A link between impaired inhibition and gamma and beta auditory entrainment. *ReScience*, 3(1), 2017.

[29] Christoph Metzner, Achim Schweikard, and Bartosz Zurowski. Multifactorial modeling of impairment of evoked gamma range oscillations in schizophrenia. *Frontiers in Computational Neuroscience*, 10, 2016.

[30] Andreas Meyer-Lindenberg and Daniel R Weinberger. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, 7(10):818–827, 2006.

[31] Robert Mill, Martin Coath, Thomas Wennekers, and Susan L Denham. A neurocomputational model of stimulus-specific adaptation to oddball and markov sequences. *PLoS computational biology*, 7(8):e1002117, 2011.

[32] P Read Montague, Raymond J Dolan, Karl J Friston, and Peter Dayan. Computational psychiatry. *Trends in cognitive sciences*, 16(1):72–80, 2012.

[33] Karen A Moxon, Greg A Gerhardt, and Lawrence E Adler. Dopaminergic modulation of the p50 auditory-evoked potential in a computer model of the ca3 region of the hippocampus: its relationship to sensory gating in schizophrenia. *Biological cybernetics*, 88(4):265–275, 2003.

[34] C Mulert, V Kirsch, Roberto Pascual-Marqui, Robert W McCarley, and Kevin M Spencer. Long-range synchrony of gamma oscillations and auditory hallucination symptoms in schizophrenia. *International Journal of Psychophysiology*, 79(1):55–63, 2011.

[35] Israel Nelken. Stimulus-specific adaptation and deviance detection in the auditory system: experiments and models. *Biological cybernetics*, 108(5):655–663, 2014.

[36] Israel Nelken and Nachum Ulanovsky. Mismatch negativity and stimulus-specific adaptation in animal models. *Journal of Psychophysiology*, 21(3-4):214–223, 2007.

[37] Brian F O'donnell, William P Hetrick, Jenifer L Vohs, Giri P Krishnan, Christine A Carroll, and Anantha Shekhar. Neural synchronization deficits to auditory stimulation in bipolar disorder. *Neuroreport*, 15(8):1369–1372, 2004.

[38] Cyrus Omar, Jonathan Aldrich, and Richard C Gerkin. Collaborative infrastructure for test-driven scientific model validation. In *Companion Proceedings of the 36th International Conference on Software Engineering*, pages 524–527. ACM, 2014.

[39] MN O'connell, A Barczak, D Ross, T McGinnis, CE Schroeder, and P Lakatos. Multi-scale entrainment of coupled neuronal oscillations in primary auditory cortex. *Frontiers in human neuroscience*, 9, 2015.

[40] RH Perlis. Translating biomarkers to clinical practice. *Molecular psychiatry*, 16(11):1076–1087, 2011.

[41] Danijela Piskulic, James S Olver, Trevor R Norman, and Paul Maruff. Behavioural studies of spatial working memory dysfunction in schizophrenia: a quantitative literature review. *Psychiatry research*, 150(2):111–121, 2007.

[42] David Fernando Ramirez-Moreno and Terrence Joseph Sejnowski. A computational model for the modulation of the prepulse inhibition of the acoustic startle reflex. *Biological cybernetics*, 106(3):169–176, 2012.

[43] Olga Rass, Giri Krishnan, Colleen A Brenner, William P Hetrick, Colleen C Merrill, Anantha Shekhar, and Brian F O'Donnell. Auditory steady state response in bipolar disorder: relation to clinical state, cognitive performance, medication status, and substance disorders. *Bipolar disorders*, 12(8):793–803, 2010.

[44] Anthony J Rissling and Gregory A Light. Neurophysiological measures of sensory registration, stimulus discrimination, and selection in schizophrenia patients. In *Behavioral neurobiology of schizophrenia and its treatment*, pages 283–309. Springer, 2010.

[45] Timm Rosburg, Nash N Boutros, and Judith M Ford. Reduced auditory evoked potential component n100 in schizophrenia—a critical review. *Psychiatry research*, 161(3):259–274, 2008.

[46] Gopal P Sarma, Travis W Jacobs, Mark D Watts, S Vahid Ghayoomie, Stephen D Larson, and Richard C Gerkin. Unit testing, model validation, and biological simulation. *F1000Research*, 5, 2016.

[47] Nestor A Schmajuk and José A Larrauri. Neural network model of prepulse inhibition. *Behavioral Neuroscience*, 119(6):1546, 2005.

[48] Wei-Xing Shi. The auditory cortex in schizophrenia. *Biological psychiatry*, 61(7):829, 2007.

[49] Peter J Siekmeier. Computational modeling of psychiatric illnesses via well-defined neurophysiological and neurocognitive biomarkers. *Neuroscience & Biobehavioral Reviews*, 57:365–380, 2015.

[50] Ray Singh and Chris Eliasmith. Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *Journal of Neuroscience*, 26(14):3667–3678, 2006.

[51] Kevin M Spencer. Baseline gamma power during auditory steady-state stimulation in schizophrenia. *Frontiers in human neuroscience*, 5:190, 2012.

[52] Kevin M Spencer, Margaret A Niznikiewicz, Paul G Nestor, Martha E Shenton, and Robert W McCarley. Left auditory cortex gamma synchronization and auditory hallucination symptoms in schizophrenia. *BMC neuroscience*, 10(1):85, 2009.

[53] Kevin M Spencer, Dean F Salisbury, Martha E Shenton, and Robert W McCarley. γ-band auditory steady-state responses are impaired in first episode psychosis. *Biological psychiatry*, 64(5):369–375, 2008.

[54] Klaas Enno Stephan and Christoph Mathys. Computational approaches to psychiatry. *Current opinion in neurobiology*, 25:85–92, 2014.

[55] Daniel Umbricht, René Koller, Liselotte Schmid, Anja Skrabo, Claudia Grübel, Theo Huber, and Hans Stassen. How specific are deficits in mismatch negativity generation to schizophrenia? *Biological psychiatry*, 53(12):1120–1131, 2003.

[56] Erricos Ventouras, Nikolaos K Uzunoglu, Dimitris Koutsouris, Charalambos Papageorgiou, A Rabavilas, and C Stefanis. Simulated generation of evoked potentials components using networks with distinct excitatory and inhibitory neurons. *IEEE Transactions on Information Technology in Biomedicine*, 4(3):238–246, 2000.

[57] Dorea Vierling-Claassen, Peter Siekmeier, Steven Stufflebeam, and Nancy Kopell. Modeling gaba alterations in schizophrenia: a link between impaired inhibition and altered gamma and beta range auditory entrainment. *Journal of Neurophysiology*, 99(5):2656–2671, 2008.

[58] Wolfger von der Behrens, Peter Bäuerle, Manfred Kössl, and Bernhard H Gaese. Correlating stimulus-specific adaptation of cortical neurons and local field potentials in the awake rat. *Journal of Neuroscience*, 29(44):13837–13849, 2009.

[59] X-J Wang, Jesper Tegnér, C Constantinidis, and PS Goldman-Rakic. Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 101(5):1368–1373, 2004.

[60] Xiao-Jing Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463, 2001.

[61] Xiao-Jing Wang and John H Krystal. Computational psychiatry. *Neuron*, 84(3):638–654, 2014.

[62] Tony W Wilson, Olivia O Hernandez, Ryan M Asherin, Peter D Teale, Martin L Reite, and Donald C Rojas. Cortical gamma generators suggest abnormal auditory circuitry in early-onset psychosis. *Cerebral Cortex*, 18(2):371–378, 2007.

[63] J Zhang, L Ma, W Li, P Yang, and L Qin. Cholinergic modulation of auditory steady-state response in the auditory cortex of the freely moving rat. *Neuroscience*, 324:29–39, 2016.