Automatically Finding Bugs in a Commercial Cyber-Physical System Development Tool Chain With SLforge

Shafiu Azam Chowdhury  
Computer Science & Eng. Dept.,  
University of Texas at Arlington, Arlington, TX, USA

Soumik Mohian  
Computer Science & Eng. Dept.,  
University of Texas at Arlington, Arlington, TX, USA

Sidharth Mehra  
Computer Science & Eng. Dept.,  
University of Texas at Arlington, Arlington, TX, USA

Siddhant Gawsane  
Computer Science & Eng. Dept.,  
University of Texas at Arlington, Arlington, TX, USA

Taylor T. Johnson  
EECS Department,  
Vanderbilt University, Nashville, TN, USA

Christoph Csallner  
Computer Science & Eng. Dept.,  
University of Texas at Arlington, Arlington, TX, USA

ABSTRACT

Cyber-physical system (CPS) development tool chains are widely used in the design, simulation, and verification of CPS data-flow models. Commercial CPS tool chains such as MathWorks’ Simulink generate artifacts such as code binaries that are widely deployed in embedded systems. Hardening such tool chains by testing is crucial since formally verifying them is currently infeasible. Existing differential testing frameworks such as CyFuzz can not generate models rich in language features, partly because these tool chains do not leverage the available informal Simulink specifications. Furthermore, no study of existing Simulink models is available, which could guide CyFuzz to generate realistic models.

To address these shortcomings, we created the first large collection of public Simulink models and used the collected models’ properties to guide random model generation. To further guide model generation we systematically collected semi-formal Simulink specifications. In our experiments on several hundred models, the resulting SLforge generator was more effective and efficient than the state-of-the-art tool CyFuzz. SLforge also found 8 new confirmed bugs in Simulink.

CCS CONCEPTS

- Software and its engineering → Model-driven software engineering; Software testing and debugging;

KEYWORDS

Cyber-physical systems, differential testing, tool chain bugs

1 INTRODUCTION

Cyber-physical system developers rely heavily on complex development environments or tool chains, which they use to design graphical models (i.e., block-diagrams) of cyber-physical systems. Such models enable engineers to do rapid prototyping of their systems through simulation and code generation [23]. Since automatically generated native code from these data-flow models are often deployed in safety-critical environments, it is crucial to eliminate bugs from cyber-physical system tool chains [7, 40].

Ideally, one should formally verify such tool chains, since a tool chain bug may compromise the fidelity of simulation results or introduce subtle bugs in generated code [8]. However, a commercial cyber-physical system (CPS) development tool chain consists of millions of lines of code, so formal verification does not (yet) scale to such tool chains. While compilers and other CPS tool chain components remain mostly unverified, we continue to observe frequent safety recalls in various industries [1, 55, 56]. The recalls are attributed to hidden bugs in the deployed CPS artifacts themselves, in spite of spending significant efforts in their design validation and verification [5, 57].

Testing, on the other hand, is a proven approach to effectively discover defects in complex software tool chains [34]. Especially randomized differential testing has recently found over a thousand bugs in popular production-grade compilers (e.g., GCC and LLVM) that are part of CPS development tool chains [17, 25, 30, 45, 58]. The technique eliminates the need of a test-oracle and can hammer the system under test in the absence of a complete formal specification of the system under test—a phenomenon we commonly observe in commercial CPS tool chain testing [6, 24, 50]. Differential testing seems suitable for black-box testing of the entire CPS tool chain, and its most susceptible parts (e.g., code generators) in particular [43, 50]. CyFuzz is the first (and only) known randomized differential testing tool for CPS data-flow languages [11].

While CyFuzz initiated the work for testing CPS tool chains, more work is necessary to evaluate the scheme’s capabilities, e.g., for finding bugs in Simulink that developers care about. For instance, a random model generator should generate tests with properties similar to the models people typically use, since they are more likely to get fixed by the system under test (SUT) developers. While large repositories of publicly available programs of various procedural and object-oriented programming languages exist [12, 22, 59], we
are not aware of such a collection of CPS models. The existing CPS studies rely on a handful of public [41] or proprietary [47] Simulink models.

Among other shortcomings, the models CyFuzz generates are small and lack many syntactic constructs. Recent studies identified expressive test-input generation as a success-factor for compiler validation ([30, 58]). Perhaps due to its inability to generate large tests with rich language features, CyFuzz has not found previously unknown bugs. Furthermore, CyFuzz essentially generates invalid models and iteratively fixes them until the SUT can compile and simulate them without error or exception. However, this heuristic approach required several time-consuming iterations and did not use Simulink specifications, which are available publicly in natural language.

To address these shortcomings, we have conducted the first study of a large number of public Simulink models. The size of many of these models is larger than the average size of models used in industry. From the collected models we obtain properties that are useful for targeting a random Simulink model generator. Our model collection is publicly available and may eliminate the nontrivial overhead of artifact-collection in future studies.

Next, extending CyFuzz, we present SLforge, a tool for automatically generating models with advanced Simulink language features. The goal is that the SLforge-generated models are similar to the collected public models. Improving on CyFuzz’s undirected random model generation approach, SLforge can generate models more efficiently, by consulting available (informal) Simulink specifications.

Finally, we provide the first approach to Equivalent modulo input (EMI) testing in CPS development tool testing [30]. SLforge creates EMI variants from the random models it generates and uses them in the differential testing setup. During an approximately five months long testing time, we found and reported 12 bugs overall, many of which were previously unknown. To summarize, the paper makes the following major contributions.

• To better target a random CPS model generator, we conduct the first large-scale study of publicly available Simulink models. A significant portion of these models are of size and complexity that are comparable to models used in industry.

• We identify problems in the existing CPS random model generator CyFuzz and design solutions that directly led to the discovery of new bugs in the Simulink tool chain.

• Finally, by comparing it with CyFuzz, we evaluate SLforge’s efficiency and bug-finding capability.

2 BACKGROUND

This section provides necessary background information on CPS dataflow models, the major commercial CPS tool-chain Simulink, the state-of-the-art differential CPS tool-chain testing tool CyFuzz, and EMI-based differential testing.

2.1 CPS Data-flow Models And Simulink

While in-depth descriptions are available elsewhere [54], the following are the key concepts. In a CPS development tool (e.g., Simulink), a user designs a CPS as a set of dataflow models. A model contains blocks. A block accepts data through its input ports, typically performs on the data some operation, and may pass output through its output ports to other blocks, along connection lines. Simulink specifies which port (of a block) supports which data-types.

In a connection, we name the block sending output source and the block receiving data a target. Output ports are numbered starting with 1. Input port numbering starts with 0, where 0 denotes a special port (e.g., the Action port of the If Action block). In addition to such explicit connections, using From and Goto blocks, one can define implicit (hidden) connections [40, 44].

Commercial CPS tool chains offer many libraries of built-in blocks. Besides creating a CPS from built-in blocks, one can add custom blocks and define their functionality via custom “native” code (e.g., in Matlab or C, using Simulink’s S-function feature). Most blocks have user-configurable parameters.

More formally, block $b \in B$ and connection $c \in C$ may be part of model $M \in M$. Then a flat (non-hierarchical) model is a tuple $(B, C)$ where $m.B$ and $m.C$ denote the model’s blocks and connections. Each connection is a tuple $(b_s, p_s, b_t, p_t)$ of source block $b_s$, source output port $p_s$, target block $b_t$, and target input port $p_t$. While a Simulink connection may have multiple targets, we break such a multi-target connection into multiple (single-target) connection tuples, without losing expressiveness.

For hierarchical models we list a model $m_i$ at hierarchy level $i$ with its $n$ direct child models as $m_i[m_i^1, \ldots, m_i^{n-1}]$. The Figure 1 example $m_1[m_2, m_3]$ has $m_1$ as its top-level model. $m_2$ and $m_3$ are $m_1$’s child models at hierarchy level 2. The dashed arrow starting at $b_2$ indicates that in the $m_2$ model $b_2$ is a placeholder for the $m_3$ model. Block $b_1$ sends data to $m_2$, where $b_0$ receives it. Block $b_0$ sends data back to $b_4$ in $m_1$.

Example placeholders are the subsystem and model reference blocks. A child model $m$’s semantics are influenced by $m$’s hierarchy-type property $T_h$, which depends on $m$’s configuration, the presence of specific blocks in $m$, and on $m$’s placeholder block-type.

After designing a model in Simulink, users typically compile and simulate it. In simulation, Simulink numerically solves the model’s mathematical relationships established by the blocks and their connections and calculates various non-dead (§4.3) block’s outputs according to user-requested sample-times and inferred context-dependent time steps, using built-in solvers [33]. Simulink offers different simulation modes. While in Normal mode Simulink “only” simulates blocks, it also emits some code for blocks in Accelerator mode, and a standalone executable in Rapid Accelerator mode.

An input port $p$ of block $b$ is Direct Feed-through if $b$’s output depends on values received from $p$. The parent to child model relation
is acyclic. But within a model Simulink permits feedback loops (circular data flow). During compilation Simulink may reject a model if it fails to numerically solve feedback loops (aka algebraic loops). Simulink also supports explicit control-flow, e.g., via if blocks. An if (“driver”) block connects to two if Action subsystem blocks, one for each branch, via the if Action block’s Action port.

2.2 Testing Simulink With CyFuzz
CyFuzz is the first known differential testing framework for CPS tool chains [11]. The framework has five phases. The first three phases create a random model and the last two phases use the model to automatically test a SUT.

Specifically, starting from an empty model, (1) the Select Blocks phase chooses random blocks and places them in the model. (2) The Connect Ports phase connects the blocks’ ports arbitrarily, yielding a model the SUT may reject, e.g., due to a type error. For example, an output port’s data type may be incompatible with the data type of the input port it connects to. (3) CyFuzz iteratively fixes such bugs in the Fix Errors phase, by responding to the SUT’s error messages with corresponding repair actions. This “feedback-driven model generation” approach, despite being an imperfect heuristic, can fix many such model errors.

Once the SUT can compile a randomly generated model, (4) CyFuzz’s Log Signals phase simulates the model under varying SUT options. The key idea of differential testing is that each such simulation is expected to produce the same results. This phase records the output data (aka signals) of each block at various time-steps. CyFuzz uses different Simulink simulation modes, partly to exercise various code generators in the tool chain. Finally, in addition to SUT crashes, (5) the Compare phase looks for signals that differ between two simulation setups, which also indicate a bug.

CyFuzz categorizes its generated models into three groups: (1) Success: models without any compile-time and runtime errors, (2) Error: models with such errors, and (3) Timed-out: models whose simulation did not complete within a configured time-out value. Although CyFuzz pioneered the differential testing of Simulink using randomly generated models, it did not find new bugs, perhaps since the generated models are small and simple (using only four built-in libraries and lacking advanced modeling features). Also, CyFuzz does not use Simulink specifications and solely relies on iterative model correction.

2.3 EMI-based Compiler Testing
Equivalent modulo input (EMI) testing is a recent advancement in differential testing of compilers for procedural languages [30]. Complementing plain differential testing, EMI found over one hundred bugs in GCC and LLVM [8]. The idea is to systematically mutate a source program as long as its semantics remain equivalent under the given input data. Engineering basic mutators is relatively easy and the overall scheme can effectively find bugs, when combined with a powerful random generator that can create expressive test inputs (e.g., Csmith [58]).

In its original implementation, EMI mainly leverages Csmith, which generates random C programs that do not take user inputs. A given compiler in a given configuration can then be expected to produce programs that yield the same result on all EMI-mutants of a given source program. The initial implementation proved very effective and found 147 bugs in production-grade C compilers such as GCC and LLVM.

3 PUBLIC SIMULINK MODEL COLLECTION
To understand the properties of CPS data-flow models designed by both researchers and engineers, we conducted the first large study of Simulink models. The largest earlier Simulink model collection we are aware of contains some 100k blocks [16]. However, these models are company-internal and thus not available for third-party studies. In contrast, our collection consists of some 145k blocks, which are all publicly available (some require a standard Simulink license as they are shipped with Simulink).

For context, earlier work reports that at Delphi, a large industrial Simulink user, an average Simulink model consists of several hundred blocks [32]. Of the models we collected, 35 consist of more than 1,000 blocks, which is larger than an average model at Delphi.

3.1 Model Collection and Classification
For this study, we used the Simulink configuration our organization has licensed, which includes the base Simulink tool chain and a large set of libraries. This configuration includes the latest Simulink version; the project web page lists the detailed configuration [10]. However, with this configuration, we could directly compile only just over half of the collected models, as the remaining ones required additional libraries that were not part of our configuration [10].

3.1.1 Tutorial. This group consists of official Simulink tutorial models from MathWorks1. We manually reviewed the models and their descriptions in the Automotive, Aerospace, Industrial Automation, General Applications, and Modeling Features categories and excluded what we considered toy examples. We also included here the Simulink-provided domain-specific library we had access to, i.e., Aerospace. An example model from this group is NASA HL-20 (1,665 blocks), which models ”the airframe of a NASA HL-20 lifting body, a low-cost complement to the Space Shuttle orbiter” [53].

3.1.2 Simple and Advanced. We collected models from both major open source project hosting services for Simulink, GitHub and Matlab Central. (1) We used the GitHub search page for keyword search (“Simulink”) and file extension search (Simulink extensions .mdl and .slx), (2) On Matlab Central2 we filtered results by “content type: model” and considered only those repositories with the highest average ratings (27 projects) or “most downloads” count in the last 30 days (27 projects).

To distinguish toy examples from more realistic models, we labeled the GitHub projects no user has forked or marked a favorite as Simple and the rest as Advanced. For the Matlab Central projects, we manually explored their descriptions and labeled those that demonstrate some blocks’ features or are academic assignments as Simple and the rest as Advanced.

As an example, a model from the Grid-Connected PV Array project is “a detailed model of a 100-kW array connected to a 25-kV grid via

1https://www.mathworks.com/help/simulink/examples.html
2https://www.mathworks.com/matlabcentral/fileexchange
3.2 Model Metrics

In this study, we focus on those model properties that are relevant for constraining a random model generator to models that are representative of realistic CPS models. Our Matlab-based tool we used to collect the following metrics is freely available on the project site [10]. The collected metric values are shown as box-plots with min-max whiskers.

3.2.1 Number of Blocks and Connections. Blocks and connections are the main elements of Simulink models and are counted widely [32, 39]. We have included the contents of masked blocks [54] in the parent model’s count. Next, we count the total number of blocks and connections at a particular hierarchy level up to hierarchy level 7.

Our connection-count metric does not include hidden connections. For connections with multiple target ports, we count the connections’ target ports. Perhaps not surprisingly, Simple models are smaller (and Advanced models are larger) than models of the other groups (Figures 2a and 2b), since we manually reviewed and classified them in this class.

3.2.2 Hierarchy Depth. Since industrial models are frequently organized as a hierarchy, we measured how deep these hierarchies are. We treated both subsystems and model reference blocks as adding a hierarchy level. Most of the collected models are indeed hierarchical (i.e., 362/391 models). But the median maximum hierarchy depth did not extend five across all model groups.

More surprising were the distribution of blocks and connections across hierarchy levels (Figures 3a and 3b). These numbers were rather similar across hierarchy levels. Overall, the number of blocks and connections in each hierarchy level were small, as denoted by the small median value.

3.2.3 Library Participation. This metric identifies the library each model block comes from. For example, do models mostly consist of built-in blocks or do they instead contain mostly custom blocks? If we cannot resolve a block’s library (i.e., due to Matlab API limitations), we record the block’s library as other.

Figure 4 suggests that only a small portion of the blocks are custom ("User_Defin"). Across all four groups, Ports & Subsystems and Math Operations were the two libraries used most frequently. SLforge thus supports these libraries (among others, see §4.1.1), and automatic custom block generation. We also noted a high contribution from the Signal Routing library using From and Goto blocks, which enables establishing hidden data-flow relationship (§2.1).

3.2.4 Requested Simulation Duration. This metric captures the total simulation time requested by a given model (not the actual CPU time spent in simulating it). Most of the models (except those from the Other group) used the default simulation duration value of 10 seconds (Figure 2d). Consequently, we ran simulations using this default value in our experiments, and have not experimented with other possible values yet.

4 SLFORGE

To address the shortcomings in the state-of-the-art differential testing framework for CPS tool chains, this section describes the design of SLforge. Figure 5 gives an overview of the SLforge’s seven main phases.

4.1 Gathering Semi-Formal Specifications

CyFuzz heavily relies on its Fix Errors phase to repeatedly compile and simulate a model and iteratively repair errors based on the content of Simulink-provided error messages. Instead of this time-consuming iterative process, SLforge aims at generating a valid model in the first place, if given the language specifications. Of course, the challenge (which motivated CyFuzz’s iterative process) is that there exists no complete, up to date, formal Simulink specification.

4.1.1 Design Choice. Simulink specifications are available as informal and semi-formal descriptions of Simulink behavior, mainly

<table>
<thead>
<tr>
<th>Group</th>
<th>M</th>
<th>C</th>
<th>H</th>
<th>Blocks</th>
<th>Connect.</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>41</td>
<td>40</td>
<td>40</td>
<td>10,926</td>
<td>11,541</td>
</tr>
<tr>
<td>s</td>
<td>156</td>
<td>99</td>
<td>136</td>
<td>7,187</td>
<td>7,121</td>
</tr>
<tr>
<td>a</td>
<td>167</td>
<td>66</td>
<td>165</td>
<td>118,632</td>
<td>116,608</td>
</tr>
<tr>
<td>o</td>
<td>28</td>
<td>14</td>
<td>21</td>
<td>8,317</td>
<td>9,577</td>
</tr>
<tr>
<td>Total</td>
<td>391</td>
<td>219</td>
<td>362</td>
<td>145,062</td>
<td>144,847</td>
</tr>
</tbody>
</table>

Figure 2: Collected public models: Total blocks (a), connections (b), maximum hierarchy depth (c), and requested simulation duration (d).

Figure 4 suggests that only a small portion of the blocks are custom ("User_Defin"). Across all four groups, Ports & Subsystems and Math Operations were the two libraries used most frequently. SLforge thus supports these libraries (among others, see §4.1.1), and automatic custom block generation. We also noted a high contribution from the Signal Routing library using From and Goto blocks, which enables establishing hidden data-flow relationship (§2.1).

4.2 Model Metrics

In this study, we focus on those model properties that are relevant for constraining a random model generator to models that are representative of realistic CPS models. Our Matlab-based tool we used to collect the following metrics is freely available on the project site [10]. The collected metric values are shown as box-plots with min-max whiskers.
from the various Simulink web sites. From our experiments with
CyFuzz, we hypothesized that many of the iterations in the Fix
Errors phase are due to block data-type inconsistency and fixing
algebraic loops. Besides, in our CyFuzz experiment (§5.1) the most
frequent error was block sample time inconsistency. We collected
specifications to both address these issues and to enable creating
large hierarchical models (as Simulink users prefer this modeling
choice).

So far, we have collected data-type support and block-parameter
specifications for all built-in libraries. Other language specifications
(§4.2) are often block and library specific. Since collecting
the entire Simulink language specification would be overwhelming,
we collected specifications for blocks from the most-used libraries.
Concretely, SLforge supports blocks from Math Operations, Ports
and Subsystems, Discrete, Continuous, Logic and Bit Operations,
Sinks and Sources libraries. This list also covers the CyFuzz-supported
libraries and thus helps ease evaluating SLforge.

4.1.2 Collection Process. Using little engineering effort, SLforge’s
regular expression based parser parsed block data-type and param-
eter specifications for all built-in blocks. However, due to the lim-
itation of the parser and Simulink’s free-form specification style,
SLforge can only collect parts of some specification. E.g., for three
different parts of the Variable Time Delay, Discrete Filter and Delay
blocks, the Direct Feed-through property (§2.1) is described as “Yes,
of the time delay (second) input”, “Only when the leading
numerator coefficient does not equal zero” and “Yes, when you clear
Prevent direct feedthrough” respectively [54].

4.2 Use of Semi-Formal Specifications

Since complete and updated formal specifications for Simulink
are not publicly available, existing work relies on a subset of Simulink
operational specifications, which are manually crafted and possibly
outdated [6, 24]. Unlike these approaches, we explored collecting
specifications directly from official Simulink documentation auto-
matically, using easy-to-engineer parsers. Such parsers can automa-
tically update the collected specifications when new versions of
the SUT are released, given the structure of the specifications go
under minor or no change. From our experience with recent ver-
sions of Simulink specifications (R2015a-R2017a), the specifications
SLforge collects indeed had minor structural changes.

Although SLforge parses specifications automatically and stores
them using internal data-structure, for the aid of discussion, we
introduce a few notions in this section. Extending the notion of
§2.1, function $T_b$ returns a block’s Simulink block-type. For example,
for each Simulink If block, $T_b$ returns If. Next, the valid predicate
indicates if the Simulink type checker and runtime system accept
a given model as legitimate, i.e., when there are no compile or
run-time exceptions. Now we can express (part of) the Simulink
specification as a formula or specification rule. Given such a rule
$\delta \in \Delta$, we denote with $m \vdash \delta$ that model $m$ satisfies (i.e.,
complies with) the rule. We observe that a valid model satisfies all (collected)
specification rules ($\forall \delta \in \Delta : valid(m) \rightarrow m \vdash \delta$).

4.2.1 Select Blocks Phase. To support built-in libraries, SLforge
uses specifications in this phase. Specifically, SLforge-generated
models satisfy the following rules by construction. For example,
using usual set cardinality notation, Eq. (1) ensures that for each If
block, the model has two IfAction subsystem blocks (one each for
the if and else branch).

$$2 \cdot | \{ b_1 \in m. B : T_b(b_1) = \text{If} \}|$$
$$= | \{ b_2 \in m. B : T_b(b_2) = \text{IfAction} \}|$$

(1)

By parsing Simulink documentation, SLforge obtains a set $S$ of
blocks from the Sinks and Source libraries that are only valid in the
top-level model, as enforced by Eq. (2). Similarly, Eq. (3) restricts
using illegitimate blocks in non-top-level models, depending on
the hierarchy-type property of the model (§2.1), using predicate
supports. The predicate holds only when model $m$’s hierarchy-
type (first argument) allows block $b$ in $m'$, based on $b$’s block-type
(second argument).

$$\forall b \in m'. B : (i > 1) \rightarrow (T_b(b) \not\in S)$$

(2)

$$\forall b \in m'. B : (i > 1) \rightarrow \text{supports}(T_b(m'), T_b(b))$$

(3)

$$\forall b \in m. B : ((T_b(m) \in W) \land stime(b) = stime(\text{driver}(m)))$$
$$\lor (T_h(m) \not\in W \land st(T_h(b), b))$$

(4)

Eq. (4) configures each block’s sample time property stime. When
used in hierarchical model $m$ of a W-listed hierarchy-type, block
$b$’s sample time should match the sample time of the model’s driver
(§2.1). In all other cases, we use predicate $st$, which holds only when
the sample time property of block $b$ is properly configured accord-
ing to its block-type. To enforce such rules, SLforge propagates
information from parent to child models.
4.2.2 Pre-connection and Connect Ports Phases. While CyFuzz uses a random search to connect unconnected ports and relies on later phases to recover from illegal connections, SLforge adds connections correctly by construction, by satisfying the following rules.

\[
\forall c \in m.C: (T_b(c.b_s) = \text{If Action}) \land (c.p_t = 0) \rightarrow (T_b(c.b_s) = \text{If})
\]

(5)

\[
\forall c \in m.C: (T_b(c.b_s) = \text{If}) \rightarrow (T_b(c.b_s) = \text{If Action}) \land (c.p_t = 0)
\]

\[
\land \{ c_2 \in m.C : c_2.b_s = c.b_s \land c_2.p_s = c.p_s \} \mid 1
\]

(6)

Eq. (5) and Eq. (6) together specify the control-flow from an If block to its If Action blocks. Specifically, each If block output port is connected to a single If Action block Action port.

4.2.3 Analyze Model Phase. On the current model state, SLforge now removes algebraic loops and assigns data-types. Instead of querying a disjoint-set data structure every time SLforge connects two blocks to detect whether connecting them will create a cycle, we detect them later in this phase using a single graph traversal.

Listing 1: Removing possible algebraic loops from a model. color(b) denotes a block’s visit-status via do_dfs method: white=unvisited; gray and black: visited.

```java
method remove_algebraic_loops (m):
    F = new set /* stores problematic blocks */
    for each block b ∈ m.B: set WHITE as color(b) for each block b ∈ m.B:
        if color(b) = WHITE: do_dfs(m, b, F)
    for each block b in F:
        s := get affected source block for b
        get and remove affected connection between s and b
        d’ := add new Delay block in m
        connect from s to d’ and from d’ to b

    method do_dfs(m, b, F):
        set GRAY as color(b)
        for each connection c ∈ m.C where c.b_s = b:
            if color(c.b_s) = WHITE: do_dfs(m, c, F)
            else if color(c.b_s) = GRAY:
                if c.p_t = 0: add b in F else: add c.b_s in F
        set BLACK as color(b)
```

After removing possible algebraic loops, SLforge propagates data-type information, to eliminate data-type mismatches. While CyFuzz compiled a model with Simulink repeatedly in the Fix Errors phase to identify data-type inconsistencies between connected blocks, SLforge fixes such errors in linear time using a single graph traversal.

Specifically, SLforge places every block whose output data-type is initially known (Non-Direct Feed-through and blocks from Source library) in a set and starting from them, runs a depth-first search on the graph-representation of the model. Data-type information is then propagated to other blocks along the connections from blocks with known output data types, using forward propagation. E.g., consider connection c ∈ m.C and say we are currently visiting block c.b_s in the depth-first search. If the data-type at c.p_s is not supported
by `c.p1` as per specification, we add a *Data-type Conversion* block between the ports.

### 4.3 EMI-testing

As recent work suggests that EMI-testing is promising for compiler testing [30], we explored this direction for Simulink. EMI-testing for Simulink could take many forms. For example, one could extend a model with blocks and connections that remain dead under the existing inputs. As another example, one could statically remove some of the dead blocks.

In this work we infer an EMI-variant from a given randomly generated model, by removing all blocks that are dead. We approximate the set of dead blocks statically, using Simulink's *block reduction* feature [54]. This approach differs from the original EMI implementation [(30)] in the sense that we collect the dead-block information statically, while [30] dynamically collected code coverage information. We chose the static approach as it required minimal engineering effort.

In our experiments, we noted that CyFuzz connects all output ports to certain *Sink* blocks. The goal was to guarantee all blocks’ participation during simulation, which allowed to use Simulink’s *Signal Logging* feature to record every block’s outputs. Consequently, CyFuzz-generated models do not have many statically dead blocks. To let EMI-testing remove larger parts of the generated model, SLforge leaves random output ports unconnected.

### 4.4 Classification of Bugs

SLforge automatically detects Simulink crash or unresponsiveness (which we categorize as *Hang/Crash Error*) and only reports if it is reproducible using the same model. Besides crash, we discuss the types of bugs in following two directions:

#### 4.4.1 Compile Time vs. Runtime

SLforge discovers some bugs during (or before) compiling the model; we categorize these as *Compile Time* bugs. In the event of compilation error, SLforge reports a bug if the error is not expected. For example, SLforge expects no data-type inconsistency error when generating type-safe models. SLforge detects some specification-mismatch bugs even before compiling, since we call various Simulink APIs to construct a model before compiling it. During this process, SLforge reports a bug when Simulink prevents it from setting a valid block parameter (according to the specification). Lastly, SLforge detects bugs when simulating the model — which we categorize as *Runtime* bugs.

#### 4.4.2 Essential Feature

Here we discuss bugs based on the essential generator/differential testing feature that helped discovering them. We attribute *Hierarchy* to a bug if SLforge can reproduce the bug only by creating hierarchical models. Next, intuitively, SLforge attributes *Specification* to a bug when it identifies Simulink specification mismatches. Finally, like CyFuzz, SLforge identifies *Comparison* bugs by simulating a model varying SUT options (§2.2). As a special case, SLforge attributes *EMI* to a bug if some EMI-variant of a successfully simulated model does not compile, or results in comparison error when after-simulation signal data of the EMI-variant is compared with the original model or with other EMI-variants.

![Figure 6: Runtime on valid models by model size given in blocks: (a) Average runtime of model generation; (b) Average number of required iterative fixes. Solid = SLforge; dashed = SLforge without specification usage and analyses; dotted = CyFuzz.](image)

### 5 EVALUATION

In this section we pose and explore the following relevant research questions.

- **RQ1** Can SLforge generate models systematically and efficiently in contrast to CyFuzz?
- **RQ2** Can SLforge generate feature-rich, hierarchical models in contrast to CyFuzz?
- **RQ3** Can SLforge effectively test Simulink to find bugs in the popular development tool chain?

To answer these research questions, we implemented SLforge on top of the open-source CyFuzz implementation. In the evaluation, we ran SLforge on 64-bit Ubuntu 16.04 virtual machines (VM), of 4 GB RAM and 4 processor cores each. We have used two identical host machines (Intel i74790 CPU (8 cores) at 3.60 GHz; 32 GB RAM each). When measuring runtime and other performance metrics (RQ1), we ran SLforge on each of the otherwise idle host machines (one VM per host). To find bugs (RQ3) we ran up to five VMs on each of these two hosts.

#### 5.1 SLforge Generates Models More Systematically and Efficiently (RQ1)

To compare SLforge’s new phases with CyFuzz in terms of efficiency and bug-finding capabilities, we conducted three experiments, each generating 160 models. To compare with CyFuzz, the experiments used blocks from four of the CyFuzz-supported libraries (i.e. *Sources, Sinks, Discrete*, and *Constant*). In the first two experiments we used SLforge: (1) enabling specification usage and analyses in Exp.S+ and (2) disabling them in Exp.S, and (3) in the third experiment Exp.CF we used CyFuzz. Across these three experiments we kept the other generator configuration parameters constant. As time-out we chose 1,200 seconds.

We compared the average time taken to generate a valid model (i.e., from Select Blocks to Fix Errors, inclusively). We also measured the number of iterative model-correction steps (*Num. Iter.*) in the Fix Errors phase. However this metric was not available in Exp.CF. In each of these experiments we started with generating 100 blocks (on average) per model, and gradually increased the average number of blocks (by 100 in each step, using 20 models in each step), up to 800 blocks/model on average. To approximate the bug-finding capability of these three setups, we counted the total number of unique bugs.
While investigating, we observed that having specifications enabled when it does not expect any data-type inconsistency between the with the number of blocks in Exp.S-. This result indicates that APIs to remove such loops, whereas SLforge eliminates the loops (median: 2,107), and an average hierarchy depth of 7 (median: 7). (excluding the Simple models), and to CyFuzz-generated models. (Figure 6b) in Exp.S+. Moreover, this value remains almost constant (valid models) rate dropped to 2%. We hypothesize that CyFuzz is not capable of generating such large hierarchical models and reduced the maximum hierarchy depth to 3. After generating 100 models in this configuration, CyFuzz achieved a 12% Success rate. In contrast, SLforge achieved a 90% success-rate with 7 depth.

Finally, a crucial step to efficiently generate large hierarchical models is to eliminate the algebraic loops from them. Simulink also rejects simulating some models in Accelerator mode in the presence of algebraic loops, which prevents differential testing using those model. As discussed before, CyFuzz depends on Simulink’s buggy APIs to remove such loops, whereas SLforge eliminates the loops in the Analyze Model phase.

5.2 SLforge Generates Large, Feature-rich Models (RQ2)

In this experiment we compare various properties of the SLforge-generated models to the models used in our study of public models (excluding the Simple models), and to CyFuzz-generated models. First, to compare SLforge with CyFuzz, we configured SLforge and CyFuzz to generate models with hierarchy depth 7, as this value is slightly larger than the median values for all model classes. For time-out parameter we chose 800 seconds and generated 100 models using each of the tools. CyFuzz’s Success (of generating valid models) rate dropped to 2%. We hypothesize that CyFuzz is not capable of generating such large hierarchical models and reduced the maximum hierarchy depth to 3. After generating 100 models in this configuration, CyFuzz achieved a 12% Success rate. In contrast, SLforge achieved a 90% success-rate with 7 depth.

In this experiment, SLforge generated models with an average of 2,152 blocks (median: 1,776), an average of 2,544 connections (median: 2,107), and an average hierarchy depth of 7 (median: 7). Both the average and median values of these properties are larger than (but still within the same order of magnitude as) the values we observed in the collected public models. SLforge-generated models are in this sense similar to the collected public models.

5.3 SLforge Found New Bugs in Simulink (RQ3)

To answer RQ3, SLforge continuously generated models and tested Simulink for approximately five months. Throughout the experiments, configuration options for SLforge varied and became applicable once we implemented a particular feature. In all of these experiments we used the Normal and Accelerator simulation modes in the comparison framework.

We have reported almost all of the SLforge-suspected bugs to MathWorks except two cases where we had associated the bug to an implementation error. For each of the reported cases, MathWorks has indicated if it considers the case a bug. For this work we mark a report as a false positive if MathWorks considers the case a non-bug. Our rate of false-positive is low: 2/12 reports.

Table 2 summarizes all the bugs we have reported. MathWorks has confirmed 10 of our reported issues as unique bugs, of which 8 are new bugs. Following are details of a representative subset of the confirmed bugs, including SLforge-generated models we manually reduced to fit the space and aid bug-reporting. Automated test-case reduction is part of future work. These models are freely available [10].

5.3.1 Hang/Crash Error bug. When generating large hierarchical models, we noticed that Matlab’s getAlgebraicLoops API hangs and makes the entire tool chain unresponsive (TSC-02513701).
Table 2: SLforge-discovered issues and confirmed Simulink bugs:

<table>
<thead>
<tr>
<th>TSC</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>02382544</td>
<td>Simulink Block parameter specification mismatch (Constant)</td>
</tr>
<tr>
<td>02382873</td>
<td>Internal rule cannot choose data-type (Add)</td>
</tr>
<tr>
<td>02386732</td>
<td>Data-type support specification mismatch (PID Controller (2DOF))</td>
</tr>
<tr>
<td>02472993</td>
<td>Automated rate transition failure (First-order hold)</td>
</tr>
<tr>
<td>02476742</td>
<td>Block-reduction optimization does not work (Accelerator mode)</td>
</tr>
<tr>
<td>02513701</td>
<td>Simulink hangs for large models with hierarchy</td>
</tr>
<tr>
<td>02515280</td>
<td>Inconsistent result and ambiguous specification (SubSystemCount metric);</td>
</tr>
<tr>
<td>02565622</td>
<td>Limited support in Accelerator mode (First-order hold)</td>
</tr>
<tr>
<td>02568029</td>
<td>timer does not execute callback as expected</td>
</tr>
<tr>
<td>02614088</td>
<td>Undocumented specification (Variable Integer Delay)</td>
</tr>
<tr>
<td>02705290</td>
<td>Incorrect data-type inheritance (multiple blocks)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>St</th>
<th>P</th>
<th>F</th>
<th>Ver</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>C</td>
<td>S</td>
<td>2015a</td>
</tr>
<tr>
<td>FP</td>
<td>C</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>C</td>
<td>S</td>
<td>2015a</td>
</tr>
<tr>
<td>NB</td>
<td>R</td>
<td>S, H</td>
<td>2017a</td>
</tr>
<tr>
<td>NB</td>
<td>R</td>
<td>E, H</td>
<td>2017a</td>
</tr>
<tr>
<td>NB</td>
<td>C</td>
<td>A, H</td>
<td>2015a</td>
</tr>
<tr>
<td>NB</td>
<td>C</td>
<td>S, H</td>
<td>2017a</td>
</tr>
<tr>
<td>KB</td>
<td>R</td>
<td>C, H</td>
<td>2015a</td>
</tr>
<tr>
<td>KB</td>
<td>C</td>
<td>S</td>
<td>2017a</td>
</tr>
<tr>
<td>NB</td>
<td>C</td>
<td>S</td>
<td>2017a</td>
</tr>
<tr>
<td>NB</td>
<td>C</td>
<td>S</td>
<td>2017a</td>
</tr>
<tr>
<td>NB</td>
<td>C</td>
<td>S</td>
<td>2017a</td>
</tr>
</tbody>
</table>

5.3.2 Specification bug. After incorporating Simulink specifications into various SLforge phases, SLforge started to identify bugs caused by specification violation. For example, bug TSC-02472993 of Figure 7 manifests when Simulink fails to handle blocks operating at different sample times, leading to runtime failure which is not expected as per specification. The bug only occurs for the First-Order Hold block and when SLforge generates hierarchical models. As another example, Figure 8 depicts Simulink’s PID Controller (2DOF) block accepting data of type unsigned int, whereas the specification states that the block only accepts data of type double (TSC-02386732).

In another case we noted that in spite of supporting type double in port d, block Variable Integer Delay (block VID in Figure 9) resulted in type-mismatch error. After reporting the issue, MathWorks suggested that the port “prefers” integer types and thus issued a type mismatch error when it was given a double type. This specification is not publicly available. Lastly, the Figure 10 issue (TSC-02515280) MathWorks classified as expected behavior, where Simulink’s count of the number of Subsystems did not match our count. However, part of Simulink’s results are inconsistent and the specification has been found ambiguous, resulting in a new confirmed bug.

5.3.3 Comparison bug. In issue TSC-02565622, one Simulink instance could simulate the SLforge-generated hierarchical model in Normal mode but returned an error in Accelerator mode, due to inconsistent block sample rates. MathWorks confirmed this as a known issue that does not have a public bug report.

5.3.4 EMI bug. Figure 12 illustrates Simulink bug TSC-02476742. Notice how only block bl11 is connected to an Outport block bl12, hence all remaining child blocks are dead and can be removed in EMI-testing. While EMI-testing in Normal mode removed all dead
child nodes, EMI-testing in Accelerator mode failed to do so, which MathWorks classified as a Simulink bug.

6 DISCUSSION
This paper performed the first large-scale study on publicly available Simulink models, to collect and observe various properties from them, which can be utilized to generate random models and serve as a collection of curated artifacts. However, some of the models are quite simple. We endeavored to classify such models in the Simple category manually, however, our approach may be imperfect and may suffer from human error. Oportunistically, we found complex and large models in our study and consequently, our collection of artifacts should be suitable for other empirical studies.

7 RELATED WORK
Empirical studies of widely used programs date back at least to the 1970s [29] and have gained increasing interest due to the wide availability of open source programs [52]. For example, earlier work computed properties from Java programs [12, 22, 59] and used the properties to guide a random Java program generator [26].

Temporo et al. presented the qualitas corpus—a curated collection of Java programs [52]. Although similar work has been performed in other domains [9, 35, 51], we are not aware of related work in the CPS domain, which differs significantly from procedural or object-oriented languages.

Recent studies introduced measures for Simulink model modularity [15] and complexity [39], but only evaluated them on a limited number of models. In contrast we created a larger collection of 391 Simulink models. Similar to [52], our model collection may serve as a corpus for Simulink-model based empirical studies.

Recent work has found many compiler bugs using differential testing. To generate programs that are syntactically correct, many of the test generators harness the language’s context-free grammar and a pre-determined probability table from which the generator chooses grammar elements [17, 25]. To generate programs that are also well-typed, McKeeman imposes the type information directly onto the stochastic grammar the generator uses [34].

Csmith, on the other hand, uses various analysis and runtime checks to generate programs with no undefined behavior [58]. Other techniques generate well-typed programs using knowledge of the type-system of the underlying language (e.g., JCrasher for Java [14]) and using constraint-logic programming (such as the Rust typechecker fuzzer [17]).

Whereas earlier approaches target compilers of textual languages (including procedural, object-oriented, and functional ones), they do not address the challenges inherent in testing CPS tool chains [11]. CyFuzz pioneered differential testing of CPS tool chains, but the prototype for Simulink was ineffective in finding new bugs. This work addresses CyFuzz’s limitations by incorporating informal Simulink specifications in the random model generation process and generating larger models with rich language features, which led to finding new bugs.

Recent work complements randomized differential testing via EMI-testing. For example, Le et al. hammer C language compilers [30]. We harness the technique to create EMI-variants of Simulink models for the first time. Other work discusses the effectiveness of randomized differential testing and EMI for OpenCL compilers and performs a comprehensive empirical compiler testing evaluation [8, 31].

Among other works, Nguyen et al. present a runtime verification framework for CPS model analysis tools leveraging random hybrid automata generation [37, 38]. In contrast, our generator does not rely on model transformations [4], which may limit the efficiency of existing work [38]. Other testing schemes target parts of the CPS tool chain utilizing graph grammars [49, 50]. However, complete and updated formal specifications for most commercial CPS development tools are unavailable and such white-box testing in parts was found undesirable [46].

Sampath et al. discuss testing CPS model-processing tools using semantic Stateflow meta-models [46]. Unfortunately, the approach does not scale and updated specifications are unavailable, due to the rapid release cycles of commercial CPS tools [17, 49]. Febrer et al. model the data-type inferencing logic of Simulink blocks for reasoning and experimental purposes [19]. While these works focus on a small part of the entire CPS tool chain, we differentially-test the entire CPS tool chain harnessing the available informal (but updated) specifications.

SLforge is loosely related to test case generators for existing Simulink models [7, 18, 20, 21, 33, 36, 48] and verification and formal analysis of CPS models [2, 3, 28, 32, 47, 60]. But they do not aim at finding bugs in the Simulink tool chain.

8 CONCLUSIONS
This paper described the first large collection of public Simulink models and used the collected models’ properties to guide random model generation. To further guide model generation we systematically collected semi-formal Simulink specifications. In our experiments on several hundred models, the resulting random Simulink model generator SLforge was more effective and efficient than the state-of-the-art tool CyFuzz. SLforge also found 8 new confirmed bugs in Simulink.

ACKNOWLEDGMENTS
The material presented in this paper is based upon work supported by the National Science Foundation (NSF) under grant numbers CNS 1464311, CNS 1713253, EPCN 1509804, SHF 1527398, and SHF 1736323, the Air Force Research Laboratory (AFRL) through the AFRL’s Visiting Faculty Research Program (VFRP) under contract number FA8750-15-1-0105, and FA8650-12-3-7255 via subcontract number WBSC 7255 SOI VU 0001, and the Air Force Office of Scientific Research (AFOSR) through AFOSR’s Summer Faculty Fellowship Program (SFFP) under contract number FA9550-15-F-0001, as well as under contract numbers FA8750-15-1-0105, and FA8650-12-3-7255 via subcontract number WBSC 7255 SOI VU 0001, and the Air Force Office of Scientific Research (AFOSR) through AFOSR’s Summer Faculty Fellowship Program (SFFP) under contract number FA9550-15-F-0001, as well as under contract numbers FA9550-15-1-0258 and FA9550-16-1-0246. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of AFRL, AFOSR, or NSF.
REFERENCES


[18]大众找茬工具成功地在Simulink中发现了Bug。


