The Thesis Committee for Marko Vasic
certifies that this is the approved version of the following thesis:

***CRN*++: Molecular Programming Language**

APPROVED BY

SUPERVISING COMMITTEE:

Sarfraz Khurshid, Supervisor

David Soloveichik, Co-Supervisor

CRN++: Molecular Programming Language

by

Marko Vasic

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ENGINEERING

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2018

To my parents, who no matter how bad moments we passed through, or how little we had, provided the crucial things needed to grow up with the 'right' set of values: a sense of warm home, support and love.

Acknowledgments

I would like to thank to my supervisor Professor Sarfraz Khurshid for his guidance, brilliant ideas, shared excitement about the project, and having trust in me. Special thanks also go to my co-supervisor Professor David Soloveichik, who introduced me to the exciting area of Molecular Programming, inspired me to pursuit it further, and provided invaluable feedback throughout the work.

In addition to my thesis supervisors, I would also like to thank Milos Gligoric for his feedback and guidance. My lab-mates and colleagues helped me make this project successful. I would like to thank Keenan Breik, Cameron Chalk, Aleksandar Milicevic, Boya Wang, and Kaiyuan Wang for sharing ideas and giving useful feedback throughout the project. Thank you Cameron Chalk for suggesting such an exciting name for the language!

The following research paper based on this work is under submission for peer-review: Marko Vasic, David Soloveichik, Sarfraz Khurshid. *CRN++*: Molecular Programming Language. Under submission in 24th International Conference On DNA Computing and Molecular Programming.

This work was funded in part by the US National Science Foundation (Grant Nos. CCF-1618895, CCF-165282, and CCF-1718903).

CRN++: Molecular Programming Language

Marko Vasic, M.S.E.
The University of Texas at Austin, 2018

Supervisors: Sarfraz Khurshid
David Soloveichik

Synthetic biology is a rapidly emerging research area, with expected wide-ranging impact in biology, nanofabrication, and medicine. A key technical challenge lies in embedding computation in molecular contexts where electronic micro-controllers cannot be inserted. This necessitates effective representation of computation using molecular components. While previous work established the Turing-completeness of chemical reactions, defining representations that are faithful, efficient, and practical remains challenging. This work introduces *CRN*++, a new language for programming deterministic (mass-action) chemical kinetics to perform computation. We present its syntax and semantics, and build a compiler translating *CRN*++ programs into chemical reactions, thereby laying the foundation of a comprehensive framework for molecular programming. Our language addresses the key challenge of embedding familiar imperative constructs into a set of chemical reactions happening simultaneously and manipulating real-valued concentrations. Although some deviation from ideal output value cannot be avoided, we develop methods to

minimize the error, and implement error analysis tools. We demonstrate the feasibility of using *CRN++* on a suite of well-known algorithms for discrete and real-valued computation. *CRN++* can be easily extended to support new commands or chemical reaction implementations, and thus provides a foundation for developing more robust and practical molecular programs.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

A highly desired goal of synthetic biology is realizing a programmable chemical controller that can operate in molecular contexts incompatible with traditional electronics. In the same way that programming electronic computers is more convenient at a higher level of abstraction than that of individual flip-flops and logic circuits, we similarly expect molecular computation to admit specification via programming languages sufficiently abstracted from the hardware. This work focuses on developing a compiler for a natural imperative programming language to a deterministic (mass-action) chemical reaction network implementing the desired algorithm. We do not directly make assumptions on how the resulting reactions would be implemented in chemistry. This could in principle be achieved by DNA strand displacement cascades [25], or other programmable chemical technologies.

Deterministic (mass-action) chemical kinetics is Turing universal [9], thus in principle allowing the implementation of arbitrary programs in chemistry. Turing universality was demonstrated by showing that arbitrary computation can be embedded in a class of polynomial ODEs [2], and then implementing these polynomial ODEs with mass-action chemical kinetics. While

these results establish a sound theoretical foundation and show the power of chemistry for handling computation tasks in general, translating and performing specific computational tasks can lead to infeasibly large and complex sets of chemical reactions.

In this work we develop a programming paradigm for chemistry, based on the familiar imperative programming languages, with the aim of making molecular programming more intuitive, and efficient. Most commonly used programming languages like Java, Python, C, etc, are imperative in that they use statements that change a program's state, with typical branching constructs such as if/else, loops, etc. Note that although CRNs are sometimes talked about as a programming language [6], they are difficult to program directly (it is even unfair to equate them with assembly language). In contrast, *CRN++* operates at a much higher level.

A mapping of imperative program logic to chemical reactions manipulating continuous concentrations poses various challenges that we must address. All reactions happen concurrently, making it difficult to represent sequential computation where, for example, the result of one operation is first computed and then used in another operation. Similarly, all branches of the program execution (i.e., if / else) are followed simultaneously to some degree.

We introduce the syntax and semantics of *CRN++*, which is, to our knowledge, the first imperative programming language which compiles to deterministic (mass-action) chemical reaction networks. *CRN++* has an extensible toolset including error analysis, as well as a simulation framework [7]. We

thus provide an automatic environment for simulating experiments based on *CRN++* programs.

A user specifies a *sequence* of statements, termed commands, to execute. Assignment, comparison, loops, conditional execution, and arithmetic operations are supported. The generated reactions are logically grouped into modules performing the corresponding command. Each module transforms initial species concentrations to their steady-state values which are the output of the module. We ensure that such modules are composable by preserving the input concentrations at the steady-state. Note that in mass-action chemistry all species occur with non-zero concentrations, and thus all reactions happen in parallel to some extent. To mimic sequential execution, we ensure that the reaction corresponding to the current command happens quickly, while other reactions are slow. For this we rely on a chemical oscillator in which the *clock* species oscillate between low and high concentrations, and sequential execution is achieved by catalyzing reactions with different clock species. To achieve conditional execution, we further need to ensure that the reactions corresponding to the correct execution branch readily occur, while those corresponding to other branches are inhibited. Our *Cmp* module sets 'flag' species to reflect the result of comparison, which catalyzes the correct branch reactions.

Sequential execution as well as conditional branching leads to errors. Error comes from the fact that instructions (reactions) that should not execute, still do (at a smaller rate, of course). Moreover, the set of basic modules, such as addition, converge to the correct value only in the limit, thus computing

approximately in finite time. To mitigate the error, we choose the set of modules to exhibit exponential (fast) convergence, and we provide a toolkit for error analysis and detection. Our tool is able to quantify error, so a user can realize where the source of error comes from, and guide the design of more optimal *CRN++* programs.

We demonstrate the expressiveness of our language by implementing and simulating common discrete algorithms such as greatest common divisor, integer division, finding integer square root, as well as real-valued algorithms such as computing Euler's number and computing π. We implement the *CRN++* compiler to reactions in Mathematica, and make use of the *CRNSimulator* package [7] for manipulation and simulation of chemical reactions. *CRN++* is an extensible programming language allowing for easy addition of new modules; we are working on the open-source version of the tool to enable others make use of it, and extend it further.

Chapter 2

Examples

In this section we discuss the characteristics of chemical reaction networks (CRNs) through examples. First, the overall idea of computation in CRNs is presented, followed by example programs in *CRN++*. The focus is to give a high level idea of our technique, while later sections discuss internal details.

Although historically the focus of the study of CRNs was on understanding the behavior of naturally occurring biological reaction networks, recent advancements in DNA synthesis coupled with general methods for realizing arbitrary CRNs with DNA strand displacement cascades [25] opened the path to engineering with chemical reactions. In this work we are not interested in a way to engineer the molecules implementing a reaction but focus on reaction behavior and dynamics. We abstract away molecule implementation information and denote molecular species with letters (e.g. A).

Molecular systems exhibit complex behaviors governed by chemical reactions. To give a formal notation of chemical reaction networks, consider the following system:

CRN 1 Example chemical reaction network

$$A + B \xrightarrow{1} A + B + C \tag{2.1}$$

$$C \xrightarrow{1} \emptyset \tag{2.2}$$

The CRN 1 consists of two reactions. A chemical reaction is defined with *reactants* (left side), *products* (right side), and *rate constant* which quantifies the rate at which reactants interact to produce products. To illustrate this, reaction 2.1 is composed of $reactants = \{A, B\}$, $products = \{A, B, C\}$, and rate constant $k = 1$. Since most reactions in *CRN++* have the rate constant equal to 1, from now on we drop the rate constant when writing reactions, unless it is different than 1. Note that multiple molecules of same species can be in a list of reactants (analogously for products); to handle this we use multiset notation. As an example, to describe reaction: $A + A \longrightarrow B$ we write $reactants = \{A^2\}$, upper indice 2 represents multiplicity (number of occurrences).

It may seem that a molecule of C is produced out of nothing in reaction 2.1, since the multiset of reactants is a submultiset of the products. This represents a level of abstraction where 'fuel' species that drive the reaction are abstracted away (i.e., the first reaction corresponds to $F + A + B \longrightarrow A + B + C$). Making this assumption allows us to focus on the computationally relevant species. The choice to use general (non-mass/energy preserving) CRNs is an established convention for DNA strand displacement cascades [25].

When the molecular counts of all species are large, and the solution is "well-mixed", the dynamics of the system can be described by ordinary differential equations (mass-action kinetics). Molecular concentrations are quantified by a system of ODEs, where concentration of each species is characterized by ODE of following structure:

$$\frac{d[s]}{dt} = \sum_{\forall rxn \in CRN} k(rxn) \cdot netChange(s, rxn) \cdot \prod_{\forall r \in reactants(rxn)} [r]^{m_{rxn}(r)}(t)$$

This ODE characterizes concentration of species s, written $[s]$. The right side is a sum over reactions in a CRN, where $k(rxn)$ is rate of reaction rxn, and $netChange(s, rxn)$ is a net change of molecules of s upon triggering of rxn (can be negative). Concentration of a reactant r in time is written $[r](t)$, while $m_{rxn}(r)$ is the multiplicity of reactant r in reaction rxn. To illustrate the general formula, the set of ODEs characterizing CRN 1 is:

$$\frac{d[A]}{dt} = 0, \frac{d[B]}{dt} = 0, \frac{d[C]}{dt} = [A](t) \cdot [B](t) - [C](t)$$

The concentration of species A and B is constant (derivatives zero); thus we can write $\frac{d[C]}{dt} = [A](0) \cdot [B](0) - [C](t)$. From this equality follows that $[C](t)$ is increasing when smaller than $[A](0) \cdot [B](0)$, decreasing in the opposite case, and equal to zero when $[C](t) = [A](0) \cdot [B](0)$. Thus this system has a global stable steady-state $[C] = [A](0) \cdot [B](0)$. We say that this module computes multiplication, due to the relation between initial concentrations and concentrations at the steady state.

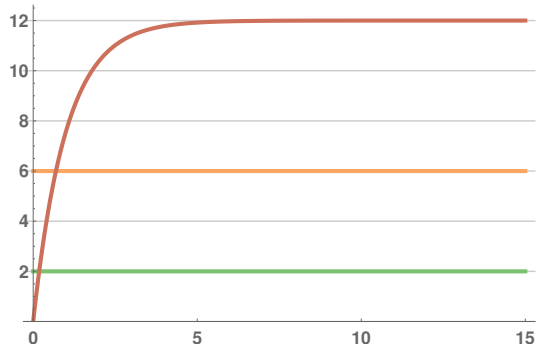


Figure 2.1: Multiplication CRN
$[A]$ shown in orange, $[B]$ in green, and $[C]$ in red.

We simulate and plot [1] dynamics of the multiplication CRN, as shown in Figure 2.1. Initial concentrations of $[A]$ and $[B]$ are 6 and 2, respectively, while the concentration of $[C]$ approaches value 12. Note that the exact value defined by the steady state ($[C](t) = 12$) is reached only at the limit of time going to infinity. Since the computation has to be done in finite time, the presence of error is unavoidable. This error raises challenging issues with programming in chemistry, and necessitates techniques for controlling it. One crucial property that determines the error is the convergence speed of the module. The multiplication command in *CRN++* is implemented through the above module, following the design principles of *convergence speed* and *composability* described in Section 3. Chemical reactions are abstracted away from the user who can simply write $Mul[A, B, C]$ to multiply.

CRN++ is imperative language, and as such supports sequential execution. Note that even a simple operation of multiplying and storing into the same variable, e.g. $A := A * B$, requires support for sequential execution: the

[1] All simulation done using *CRNSimulator* package developed by David Soloveichik [7]

above implementation of the *Mul* module necessarily assumes that the output species is different from the input species. Otherwise, $Mul[A, B, A]$ goes to infinity or 0 depending on the value of B. To implement this operation, we split the computation into two sequential steps: (1) $C := A*B$, (2) $A := C$. For the multiplication we use the *Mul* module described above. For the assignment we use the load module (*Ld*). To ensure the assignment executes after the multiplication is finished, we catalyze the two modules with the clock species that reach their high values in different phases of the oscillator. Importantly, the chemical oscillator and clock species are abstracted away from the user, who simply uses the *step* construct to order reactions. To implement the desired operation the user would write the code like in Figure 2.2.

```
1 crn={
2   step[{Mul[a,b,c]}],
3   step[{Ld[c,a]}]
4 };
```

Figure 2.2: *CRN++* program computing *A:=A*B*

One of the basic blocks of programming languages are conditional branches, executing upon success of a precondition. Similarly to implementing sequential operations, we implement conditional execution by activating (through catalysis) some reactions and deactivating others, depending on a result of condition. Since no species can be driven to 0 in finite time[2], all branches of condition will be active to some extent, which makes this an inter-

[2]Although certain pathological CRNs can drive concentrations to infinity in finite time (e.g., $2A \rightarrow 3A$), and thereby drive certain other species to 0 in finite time (e.g., with an additional $B + A \rightarrow A$), these cases cannot be implemented with any reasonable chemistry.

esting source of errors without direct analogy in digital electronics. In contrast to sequential computation catalyzed by clock species, conditional blocks are catalyzed by so-called 'flag' species. Flag species have high and low values that reflect the result of the comparison. We provide the *Cmp* module which sets the flag species to reflect the result of the comparison. The code in Figure 2.4a uses the *Cmp* module to compare a and b (step 1), enabling conditional execution in the next step.

To demonstrate the expressiveness of our language we showcase the implementation of Euclid's algorithm (Figure 2.3) to compute the greatest common divisor (GCD) of two numbers. The GCD is found by subtracting the smaller of the values from larger until the point when the two values become equal.

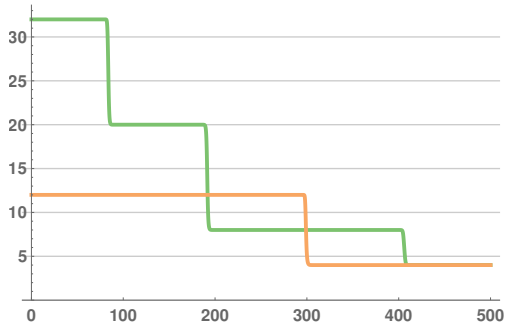
```
1: procedure GCD(a, b)
2:     while a ≠ b do
3:         if a > b then
4:             a ← a − b
5:         else
6:             b ← b − a
7:         end if
8:     end while
9:     return a
10: end procedure
```

Figure 2.3: Euclid's algorithm

Figure 2.4a shows the implementation of Euclid's algorithm in *CRN++*. Lines 2-3 define the initial concentrations of species a, b; $a0$ and $b0$ represent the values for which GCD is computed. To order the execution, the *step* construct is used. In the first step a and b are stored into temporary variables

```
crn = {
  conc[a,a0],
  conc[b,b0],
  step[{
    Ld[a, atmp],
    Ld[b, btmp],
    Cmp[a,b]
  }],
  step[{
    IfGT[{ Sub[atmp,btmp,a] }],
    IfLT[{ Sub[btmp,atmp,b] }]
  }]
};
```

(a) GCD implementation



(b) Dynamic simulation of the GCD program for $a0 = 32$, $b0 = 12$. Concentrations of a (green), and b (orange) are shown in function of time.

Figure 2.4: GCD in *CRN*++: implementation and simulation

and compared, setting the flag species to reflect the result of the comparison. The second step uses the result of the previous comparison, and effectively stores $a - b$ into a if $a > b$, and vice versa. Since the same species cannot be used as both input and output to *Sub* module, temporary variables are used (*atmp* and *btmp*). Steps repeatedly execute due to the oscillatory behavior of the clock species, thus implementing looping behavior by default; the steps can be viewed as being inside of the 'forever' loop. *CRN*++, in addition to the language and compiler to chemical reactions, is connected to the simulation backend that enables convenient testing for correctness. We show simulation of the GCD program in Figure 2.4b where GCD(32,12) is computed. Steps repeatedly trigger causing a and b to converge to the correct result after a couple iterations.

In addition, we implement a set of algorithms in (a) Discrete space—counter, integer division, integer square root, as well as in (b) Continuous space, by implementing *CRN++* programs that approximate value of *Euler's* constant and π. These examples are shown in Section 4.

Chapter 3

Technique

This section explains *CRN++*, both the underlying constructs used to build it, as well as high level primitives that represent the language itself. We start by presenting high-level modules that are at the core of *CRN++* (Section 3.1), followed by explanation about how the sequential behavior is achieved (Section 3.2), after which we give an overview of *CRN++* grammar (Section 3.3), finally we talk about error detection and analysis tools we provide (Section 3.4).

3.1 Modules

Modules represent the core of *CRN++*, and in their form are somewhat analogous to the instruction set architecture (ISA) in machine languages. Modules implement basic operations such as load, add, subtract, multiply, compare.

There are multiple ways of computing addition and other operations in chemistry. As mentioned in the previous section, our implementation choice is led by two basic principles: (a) convergence speed, (b) composability.

3.1.1 Convergence speed

Consider CRN implementing addition:

CRN 2 Addition CRN (inputs preserved). Inputs: A and B, output: C.

$$\begin{aligned} A &\longrightarrow A + C \\ B &\longrightarrow B + C \\ C &\longrightarrow \emptyset \end{aligned}$$

By solving the system of ODEs that characterize the concentration of C we get the following:

$$[C](t) = [A] + [B] + ([C](0) - [A] - [B]) \cdot e^{-t}$$

$[C](t)$ is concentration of species C in time, accordingly $[C](0)$ is initial concentration, while $[A](t)$ and $[B](t)$ are constant (not dependent on time) thus we exclude parenthesis and write $[A]$ and $[B]$. From the equation it follows that $[C]$ converges to the value $[A] + [B]$, and thus we say the CRN performs addition. To consider the convergence speed we look at the non-constant part of the equation. Due to the factor e^{-t} the decrease of the non-constant part is exponential, thus we say that the CRN exhibits *exponential* convergence speed. The convergence speed is of great importance, since it directly affects computation error; the sooner reaction converges the sooner it approaches the correct value.

3.1.2 Composability

There are alternative ways to implement addition and have exponential convergence speed:

CRN 3 Addition CRN (destructs inputs). Inputs: A and B, output: C.

$$A \longrightarrow C$$
$$B \longrightarrow C$$

Note that the above module does not preserve the input values. For easier discussion, we name the initial addition module $CRNAdd_1$ (CRN 2), and the above one $CRNAdd_2$ (CRN 3). To compute $E := (A * B) + D$ we combine the *Mul* module (CRN 1), computing $C := A * B$, with an addition module, computing $E := C + D$. Consider combining the multiplication module with one of the addition modules. If $CRNAdd_1$ is used, multiplication converges to the correct value, after which $CRNAdd_1$ has correct inputs and converges to the expected value – $E := (A * B) + D$. Before the multiplication converges, C becomes equal to $A * B$, reactions of $CRNAdd_1$ trigger, but since the module is input-preserving they do not affect steady state of the multiplication module. However, $CRNAdd_2$ consumes its inputs, and the composition will give incorrect result. The *Mul* CRN constantly drives C to value $A * B$, and will keep refilling inputs to the $CRNAdd_2$, causing the wrong result. This is the reason $CRNAdd_1$ is preferred over $CRNAdd_2$. More formal discussion of composability is presented in work by Buisman et al. [3], including proof showing the composed module has a unique stable steady state, and that it preserves

the convergence speed.

We have set up the two main design criteria (convergence speed and composability) for the modules, and we next describe the core modules of *CRN++*.

3.1.3 *Ld* Module

Loads the value from source (first argument) into a destination (second argument). The CRN used for load operation is following:

CRN 4 Load CRN

$$A \longrightarrow A + B$$
$$B \longrightarrow \emptyset$$

A are input and B are output species. This module, similar to *Add*, has exponential convergence speed (detailed analysis in Buisman et al. work [3]). In addition, concentrations of inputs are constant, thus ensuring composability.

3.1.4 *Add* Module

Adds two values (first and second argument) and stores the result into destination (third argument). The Add CRN is shown in CRN 2; its convergence speed and composability are already discussed.

3.1.5 *Sub* Module

Subtracts the second input value from the first and stores into destination (third argument).

CRN 5 Subtraction CRN

$$
\begin{aligned}
A &\longrightarrow A + C \\
B &\longrightarrow B + H \\
C &\longrightarrow \emptyset \\
C + H &\longrightarrow \emptyset
\end{aligned}
$$

The above CRN was generated via evolutionary algorithms [3]; by analyzing its system of ODEs, the network computes subtraction. Input species A and B are not affected and the property of composability is satisfied. Neither we nor Buisman et al. managed to find the analytical solution; however, analysis shows that the module converges exponentially quickly unless $A = B$ (see the Alternative Design subsection of the Cmp module below for an analogous, easy to analyze case.) In our examples, A and B usually differ by at least 1. Our error evaluation tools (Section 3.4) help in detecting and analyzing problematic cases (e.g., where A and B are close), thus enabling the user to redesign the CRN. Runtime assertions in the simulation package that automatically notify the user about these kind of problems would help identify the source of the error. Note that many algorithms can be refactored to reduce the error (see Section 6).

3.1.6 *Mul* Module

Multiplies inputs (first and second argument) and stores into destination (third argument). The multiplication CRN is shown in Section 2. This CRN does not affect inputs and has exponential convergence speed, as shown by Buisman et al. [3].

We presented modules for performing arithmetic operations (*Ld*, *Add*, *Sub*, *Mul*). The CRNs impose the restriction that same species cannot be used as both input and output. More generally, species used as input to a module cannot be used as output of any other module executing in the same step. We show the exhaustive list of modules in Table 3.1. Importantly, *CRN++* is extensible, and supports easy addition of new modules.

3.1.7 *Cmp* Module

Compares the two values, and produces signals (flag species) informing which value is greater or if they are equal.

Alternative Designs. Before explaining our implementation of comparison we discuss alternative implementations, and point out design decisions that lead to the current implementation. One of more obvious ways to implement comparison is using following reaction:

$$A + B \longrightarrow \emptyset \tag{3.1}$$

If initially $[A] > [B]$ than all molecules of B interact with A, leaving molecules of A at the equilibrium, and analogously for $[B] > [A]$. To conditionally

execute when $[A] > [B]$, one can trigger reaction 3.1 in one clock phase and than use A catalytically in the next clock phase; conditionally executing when $[B] > [A]$ is symmetrical.

The comparison module proposed above does not preserve inputs, and thus it is not composable. This imposes the restriction that in the step in which comparison is used no other module uses the compared values. Our *Cmp* module does not have this restriction.

We analyze the ODE describing this CRN to evaluate the convergence speed. Since the amount of B decreases with the same speed as A, we can express $[B](t) = [A](t) + d_0$, where $d_0 = [B](0) - [A](0)$. The following equation holds:

$$\frac{d[A]}{dt} = -[A](t) * ([A](t) + d_0) \implies [A](t) = \frac{a_0 d_0}{-a_0 + a_0 e^{d_0 t} + d_0 e^{d_0 t}}$$

If $d_0 > 0$ ($[B](0) > [A](0)$) terms with exponential factors converge to infinity, and $[A]$ to zero. Conversely, when $d_0 < 0$, exponential factors converge to zero, and $[A]$ to $-d_0$. A converges exponentially, unless A and B are equal at the beginning ($d_0 = 0$); then the dynamics of A are described with:

$$[A](t) = \frac{a_0}{1 + a_0 t}$$

In conclusion, the module converges fast (exponential speed) when operands are different, while the module converges slow (linear speed) when operands are equal (or close to each other). The linear convergence speed is yet another problem that lead to sub-optimal performance of this module. Recall that the

comparison module drives the flag species which then catalyze branches that should execute, thus having a chained effect. It is of great importance to have a reliable comparison module.

Lastly, to detect equality with the above proposed module, absence of a species needs to be detected, since both values are driven to zero in case of equality. Detecting the absence of species in chemistry is itself non-trivial and error-prone. There are several approaches based on so-called *absence indicators*. Generally speaking, the absence indicator for A is produced at a constant rate and gets degraded by A. The absence indicator has to be produced slowly, or else it will be present in non-negligible concentration even if A is present. The absence indicators in the literature rely on a difference between rate constants of several orders of magnitude. The relatively slow dynamics of the production of the absence indicator lead to a fair amount of error affecting the computation, and necessitate slowing down the clock (i.e., the whole computation) to work properly.

Our Design. *Cmp* is implemented using two sequentially executed sets of reactions, which trigger in consecutive clock phases. In the first phase, the input values (X and Y) are normalized to signal species $XGTY$ and $XLTY$ (CRN 6). For example, if $[X] = 80$ and $[Y] = 20$, signal species $XGTY$ and $XLTY$ converge to 0.8 and 0.2, respectively. We analyze the ODEs characterizing the normalization module and conclude it exhibits exponential convergence speed.

CRN 6 CRN for normalizing compared values

$$XGTY + Y \longrightarrow XLTY + Y$$
$$XLTY + X \longrightarrow XGTY + X$$

The goal of the second phase of comparison is to detect which normalized value is greater. We use a chemical *Approximate Majority* (AM) algorithm [5] to detect if $XGTY$ or $YGTX$ is in majority. All molecules of a less populous species convert to other species. AM reactions are:

CRN 7 Approximate Majority CRN

$$XGTY + XLTY \longrightarrow XLTY + B$$
$$B + XLTY \longrightarrow XLTY + XLTY$$
$$XLTY + XGTY \longrightarrow XGTY + B$$
$$B + XGTY \longrightarrow XGTY + XGTY$$

The majority algorithm causes convergence of $XGTY$ to 1 and $XLTY$ to 0 when $X > Y$, and vice versa. Now, one can use species $XGTY$ as a catalysts in reactions that should execute only if $X > Y$, and $XLTY$ if $X < Y$. The AM network has been studied in the stochastic context (stochastic CRNs) and is known to converge quickly, even when inputs are close [1].

Equality checking. Due to the always present error in chemical computation, checking for equality is actually approximate-equality checking. Consider having a chemical program with real values, then if the values are close to each other it is impossible to tell if they are actually equal but affected

with error, or they represent different real valued signals. Due to this issue, while comparing for equality is impossible, we compare for ϵ-range equality, where ϵ can be arbitrarily small. Since most of the problems we solve are discrete algorithms we use equality checking for $\epsilon = 0.5$, allowing easy comparison of the integer values (e.g., values in range $(2.5, 3.5)$ are considered to be equal to 2). To support equality checking we compare $x + \epsilon$ with y (generating signals $XGTY$ and $XLTY$), and at the same time compare $y + \epsilon$ with x (generating signals $YGTX$ and $YLTX$). Combining the signals of the two comparisons gives the desired result: If $X = Y$, signal $XGTY$ is high ($XLTY$ low) and $YGTX$ is high ($YLTX$ low) due to the added offset. To execute a reaction upon equality both $XGTY$ and $YGTX$ are used catalytically. If $X > Y$, signal $XGTY$ is high ($XLTY$ low) and $YLTX$ is high ($YGTX$ low), so both $XGTY$ and $YLTX$ should be used catalytically. Symmetrically for $X < Y$, both $XLTY$ and $YGTX$ are used catalytically. Note that unlike in the previously proposed comparison module, this module does not ask for absence checks and absence indicators, and as such is more reliable in time-constrained environment. After calling *Cmp* in a step, programmer can use *IfGT*, *IfGE*, *IfEQ*, *IfLT*, *IfLE* in subsequent steps to conditionally execute reactions. Note that the flags are active until the next call to *Cmp* module.

3.2 Sequential Execution

CRN++ allows programming in a sequential manner, despite the intrinsically parallel nature of CRNs. To model sequential execution in CRNs there

Mnemonic	Restrictions	Output (Steady State)	CRN
Ld	$B \not\equiv A$	$B := A$	$A \longrightarrow A + B$ $B \longrightarrow \emptyset$
Add	$C \not\equiv A \wedge C \neq B$	$C := A + B$	$A \longrightarrow A + C$ $B \longrightarrow B + C$ $C \longrightarrow \emptyset$
Sub	$C \not\equiv A \wedge C \not\equiv B$	$C := \begin{cases} A - B, & A > B \\ 0, & \text{otherwise} \end{cases}$	$A \longrightarrow A + C$ $B \longrightarrow B + H$ $C \longrightarrow \emptyset$ $C + H \longrightarrow \emptyset$
Mul	$C \not\equiv A \wedge C \not\equiv B$	$C := A \cdot B$	$A + B \longrightarrow A + B + C$ $C \longrightarrow \emptyset$
Dvd	$C \not\equiv A \wedge C \not\equiv B$	$C := A/B$	$A \longrightarrow A + C$ $B + C \longrightarrow B$
Sqr	$B \not\equiv A$	$B := \sqrt{A}$	$A \xrightarrow{1} A + B$ $B + B \xrightarrow{\frac{1}{2}} \emptyset$
AM	$A \not\equiv B$	$A := \begin{cases} A + B, & A > B \\ 0, & B > A \end{cases}$ $B := \begin{cases} 0, & A > B \\ A + B, & B > A \end{cases}$	$A + B \longrightarrow A + T$ $B + A \longrightarrow B + T$ $T + A \longrightarrow A + A$ $T + B \longrightarrow B + B$
Cmp	$A \not\equiv B$	Sets flag species	* Two CRNs (normalization and AM) triggering in two consecutive phases (discussed in Section 3)

Table 3.1: *CRN*++ Modules

The first column denotes the name of the module. The restrictions column imposes compile-time restrictions for using modules, here $\not\equiv$ is used to mean different species (not values). The output column shows the value of outputs at the steady state. Finally, the CRN column shows chemical reactions implementing the module.

is a need to isolate two reactions from co-occurring, and control the order in which they happen. The key construct we rely on to achieve these goals is a chemical oscillator.

A chemical oscillator is a CRN consisted of species which concentrations oscillate between low and high values. The oscillatory CRN [18] we use is described with a following set of reactions:

CRN 8 Oscillator CRN

$$\begin{aligned} i = 1, ..., n-1 : X_i + X_{i+1} &\longrightarrow 2\,X_{i+1} \\ X_n + X_1 &\longrightarrow 2\,X_1 \end{aligned}$$

X_i are clock species, and n is number of them. Concentration of X_i oscillates between zero and maximum value – which depends on initial concentrations. Catalytic addition of the clock species to reactions controls the rate at which the reaction fires. All X_i oscillate at the same frequency, but differ in oscillation (clock) phase. Different species have different oscillation phase and reach minimum and maximum points at different time moments, as shown in Figure 3.1. To ensure two reactions (rxn_1 and rxn_2) do not co-occur, we catalyze reactions with two non-overlapping clock species. It is not possible to ensure that two clock species have no overlap, and to allow for correct sequential execution it is important to keep it as low as possible. We use every third clock species, i.e. X_3, X_6, X_9 etc., to catalyze reactions that should be ordered.

The chemical oscillator is abstracted from a *CRN++* user, who can or-

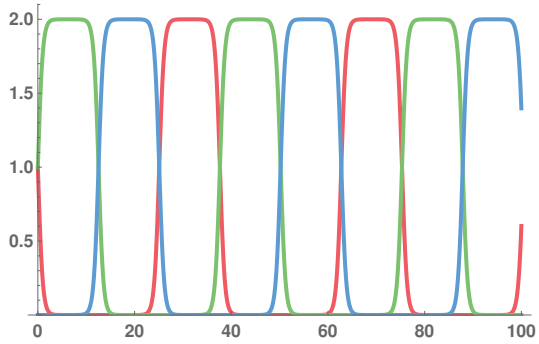


Figure 3.1: Chemical oscillator containing 3 species: X_1 (red), X_2 (green), and X_3 (blue)

der reactions using the *step* construct. Reactions in different calls of *step* are isolated from each other through clock species acting catalytically. Steps are assigned in order, meaning that earlier calls of *step* are assigned earlier phases. The oscillatory behavior of the clock species causes steps to get repeated, after the last step is executed the first one starts again, causing the loop-like behavior. The total number of clock species needed is automatically determined through the number of calls to *step*. Each call to *step* typically requires one clock species, with the exception in the case *Cmp* module is used, which requires two phases to execute.

3.3 Grammar

We already revealed many pieces of *CRN++* syntax, but have not coherently presented it. Figure 3.2 shows overview of the grammar.

At its root CRN contains a list of *RootS*s, where *RootS* can be either *ConcS* – defines initial concentration of species, *RxnS* – defines a reaction, *ArithmeticS* – performs arithmetic operation, and *StepS* – orders execution.

⟨*Crn*⟩ ::= ‘`crn= {`’ ⟨*RootSList*⟩ ‘`}`’

⟨*RootSList*⟩ ::= ⟨*RootS*⟩
| ⟨*RootS*⟩ ‘`,`’ ⟨*RootSList*⟩

⟨*RootS*⟩ ::= ⟨*ConcS*⟩
| ⟨*RxnS*⟩
| ⟨*ArithmeticS*⟩
| ⟨*StepS*⟩

⟨*ConcS*⟩ ::= ‘`conc[`’⟨*species*⟩‘`, `’⟨*number*⟩‘`]`’

⟨*RxnS*⟩ ::= ‘`rxn[`’⟨*Expr*⟩‘`,`’⟨*Expr*⟩‘`,`’⟨*number*⟩‘`]`’

⟨*ArithmeticS*⟩ ::= ‘`Ld [`’⟨*species*⟩‘`,`’⟨*species*⟩`]`
| ‘`Add [`’⟨*species*⟩‘`,`’⟨*species*⟩‘`,`’⟨*species*⟩`]`
| ‘`Sub [`’⟨*species*⟩‘`,`’⟨*species*⟩‘`,`’⟨*species*⟩`]`
| ‘`Mul [`’⟨*species*⟩‘`,`’⟨*species*⟩‘`,`’⟨*species*⟩`]`

⟨*CmpS*⟩ ::= ‘`Cmp [`’⟨*species*⟩‘`,`’⟨*species*⟩`]`

⟨*StepS*⟩ ::= ‘`step [`’ *NestedSList*‘`]`’

⟨*NestedSList*⟩ ::= ⟨*NestedS*⟩
| ⟨*NestedS*⟩ ‘`,`’ ⟨*NestedSList*⟩

⟨*NestedS*⟩ ::= ⟨*RxnS*⟩
| ⟨*ArithmeticS*⟩
| ⟨*CmpS*⟩
| ⟨*ConditionalS*⟩

⟨*ConditionalS*⟩ ::= ‘`IfPresent[`’⟨*species*⟩‘`,`’⟨*NestedSList*⟩‘`]`’
| ‘`IfGT [`’⟨*NestedSList*⟩‘`]`’
| ‘`IfGE [`’⟨*NestedSList*⟩‘`]`’
| ‘`IfEQ [`’⟨*NestedSList*⟩‘`]`’
| ‘`IfLT [`’⟨*NestedSList*⟩‘`]`’
| ‘`IfLE [`’⟨*NestedSList*⟩‘`]`’

⟨*Expr*⟩ ::= ⟨*species*⟩ { ‘`+`’ ⟨*species*⟩ }

Figure 3.2: *CRN*++ Grammar

Furthermore, *StepS* is divided into a list of *NestedS*s, where each *NestedS* is either *RxnS*, *ArithmeticS*, *CmpS* – performs comparison, or *ConditionalS*. *ConditionalS* conditionally executes a block based on result of previous comparison. Note that comparison should be executed in any step prior to conditional execution. Based on result of comparison, whether the first operand is *greater than*, *greater or equal*, *equal*, *less or equal*, *less than* the second operand, conditional block *IfGT*, *IfGE*, *IfEQ*, *IfLT*, *IfLE* is executed. To execute upon a presence of a species, *IfPresent* can be used; which catalytically adds species to reactions.

The grammar can be easily extended; e.g., new arithmetic modules can be added to the list of *ArithmeticS* nonterminals. Also, we experimented with absence indicators, *CRN++* grammar allows for easy addition of *IfAbsent* conditional statements that can be used to compare synchronous and asynchronous programs.

3.4 Error Evaluation

Programming chemistry is inherently error-prone. We identify three specific sources of error in *CRN++*. First, CRNs converge asymptotically—only in the limit is the correct value reached— thus leaving certain amount of error in a finite time. Second, we cannot completely turn off modules which are not supposed to be currently executing, whether they belong to another sequential step, or to another branch of execution. In addition, comparison has to take into account possible error in the compared values.

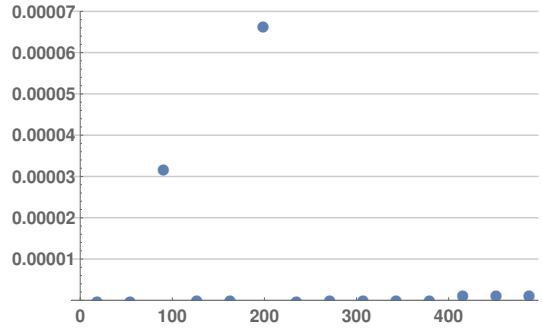


Figure 3.3: Error evaluation of species a from GCD program.

Our design decisions were based on minimizing the error; however since error cannot be avoided altogether, we provide a toolkit that helps in error analysis and guiding the CRN (program) design. Using the tool, users can, for any species of interest, track the difference between the correct value, and the (simulated) value in chemistry. For example, if operation $Add[a, b, c]$ is executed in a step, than $c = a + b$ is expected in the following step. *CRN++* allows measuring the difference between the expected $c = a + b$, and actual simulation value. This helps users analyze the error, and detect if the error builds up over time.

We analyze the value of operand a from GCD example Figure 2.4, and plot the error in Figure 3.3. In Figure 3.3, the x-axis represents time, while the y-axis shows the difference between expected and actual value of a. Note that the error is sufficiently small that the algorithm executes correctly throughout the analyzed time. The error is not constant, which opens interesting questions of correlating the error with instructions in the program. To correlate error with program instructions we examine the GCD simulation (Figure 2.4b). It is easy to connect the first two spikes of error with subtraction of a.

We provide the error evaluation framework with the vision of it being a guiding element for programming in *CRN++*. We found this technique particularly useful for validation of programs, analyzing the error, understanding the sources of error, and redesigning the CRN for correctness.

Chapter 4

Application

In this section we demonstrate *CRN++* on several examples (Section 4.1). There are two classes of problems we consider, (a) Discrete space—problems involving discrete (integer) values, such as computing the greatest common divisor, discrete counter (Section 4.1.1), factorial (Section 4.1.2), integer division (Section 4.1.3), integer square root (Section 4.1.4); (b) Continuous space—problems including real valued values such as computing the *Euler*'s (Section 4.1.5) number and the number π (Section 4.1.6). Furthermore, we talk about the error evaluation in Section 4.2.

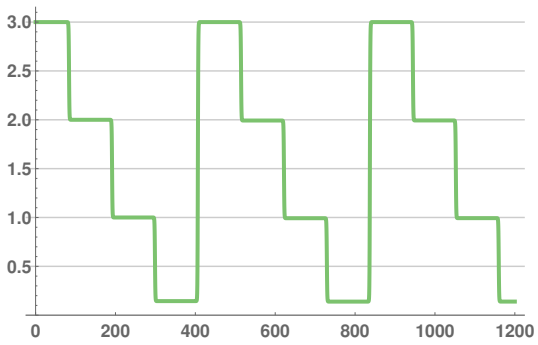
4.1 Examples

4.1.1 Discrete Counter

We implement a discrete counter that counts from a predefined value to zero, and repeats the process. Fig 4.1 shows both *CRN++* program and simulation results. The counter value is stored in the variable c, $cInitial$ preserves the initial value of the counter for later refills, while *one* and *zero* store constants 0 and 1, respectively. The initial concentrations of the species are set up in Lines 2-5, note that $c0$ is a parameter representing the initial

```
1 crn = {
2   conc[c,c0],
3   conc[cInitial,c0],
4   conc[one,1],
5   conc[zero,0],
6   step[{
7     Sub[c,one,cnext],
8     Cmp[c,zero]
9   }],
10  step[{
11    IfGT[{ Ld[cnext,c] }],
12    IfLE[{ Ld[cInitial,c] }]
13  }]
14 }
```

(a) *CRN++* code



(b) Simulation results for $c0 = 3$; value of c is shown (green line).

Figure 4.1: Discrete counter in *CRN++*

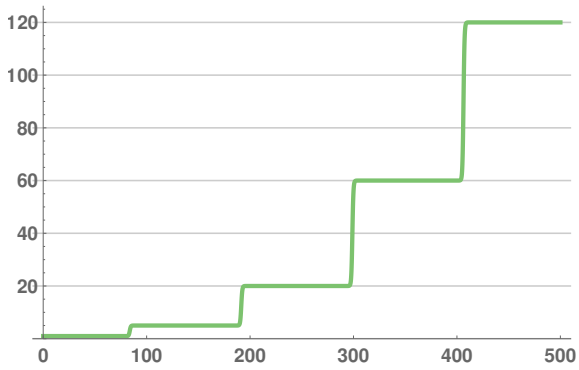
counter value. The counter is subtracted by one ($cnext := c-1$), and compared with the zero; in the first step. In a case counter is zero, than its value is reset to the initial value ($c := cInitial$), otherwise it is decremented by one ($c := cnext$), in the second step. Recall the steps exhibit looping behavior, thus the above process is repetitive.

4.1.2 Factorial

We compute the factorial using *CRN++* program. Fig 4.2 shows both the program and simulation results. To compute the factorial of a number n, we store n in the iterator variable i, and repeatitively multiply f with i, decreasing i until it gets to zero. Initial concentrations of the species are defined in Lines 2-4. In the first step (Lines 5-9), value of the iterator i is compared

```
1 crn={
2   conc[f,1],
3   conc[one,1],
4   conc[i,f0],
5   step[{
6     Cmp[i,one],
7     Mul[f,i,fnext],
8     Sub[i,one,inext]
9   }],
10  step[{
11    IfGT[{
12      Ld[inext,i],
13      Ld[fnext,f]
14    }]
15  }]
16 }
```

(a) *CRN++* code



(b) Simulation results for $f0 = 5$; value of f is shown (green line).

Figure 4.2: Factorial in *CRN++*

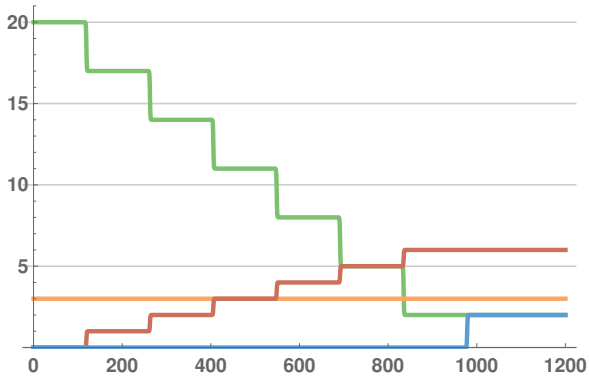
with one to check the termination condition; f is multiplied with the i storing the value in the temporary variable $fnext$, and finally the iterator is decremented storing the value in the temporary $inext$. In the second step (Lines 10-15), commands are executed only $i > 1$, moving the values of temporaries back to f, and i.

4.1.3 Integer Division

We implement integer division of a two numbers, computing quotient and remainer of the operation. Dividend is stored in the variable a, divisor in b, quotient in q, and remainder in r. Fig 4.3 shows both program and simulation results. Value of the divisor is subtracted from the dividend, until dividend becomes less than the divisor. In the first step (Lines 5-7), dividend

```
1 crn={
2   conc[a,a0],
3   conc[b,b0],
4   conc[one,1],
5   step[{
6     Cmp[a,b]
7   }],
8   step[{
9     IfGE[{
10      Sub[a,b,anext],
11      Add[q,one,qnext]
12    }]
13  }],
14  step[{
15    IfGE[{
16      Ld[anext,a],
17      Ld[qnext,q]
18    }],
19    IfLT[{Ld[a,r]}]
20  }]
21 };
```

(a) *CRN++* code



(b) Simulation results for $a0 = 20$, $b0 = 3$; values of a (green), b (orange), q (red), and of r (blue) are shown.

Figure 4.3: Integer division in *CRN++*

and divisor are compared to detect if the termination condition is satisfied. In the second step (Lines 8-13), if $a > b$, the dividend is subtracted by divisor and quotient incremented. In the third step (Lines 14-20), if $a > b$, new values for the dividend and quotient are restored from the temporary variables into the original ones. Also, in the last step, if $a \leq b$, the value of dividend is stored into the remainder (line 17).

```
1: procedure INT SQRT(n)
2:     z ← 0
3:     while (z + 1)^2 ≤ n do
4:         z ← z + 1
5:     end while
6:     return z
7: end procedure
```

Figure 4.4: Integer square root algorithm

4.1.4 Integer Square Root

We implement a program that finds integer square root of a number. Figure 4.4 shows the algorithm; the square root of a number n is found by iterating through numbers $0, 1, 2$, etc, until the power of the iterated number overshoots n. We map the algorithm to *CRN++* program, and show the code and simulation results in Figure 4.5. In the first step (Lines 3-7), we increment the z ($znext := z + 1$), compute power of the $z + 1$ ($zpow := znext * znext$), and compare the power with n. In the second step (Lines 8-11), if $zpow < n$, then $znext$ is stored into z, otherwise, the result is computed and stored in the *out*.

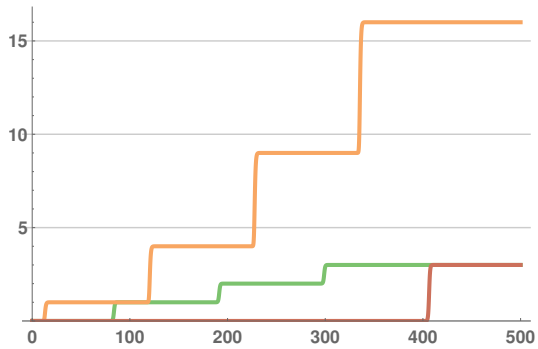
4.1.5 Euler's number approximation

So far, we presented discrete algorithms, however chemistry allows for real-valued (analog) computations. For programming with real values we extend *CRN++* with additional module performing division – *Dvd*. *Dvd* module follows same design principles and characteristics as other arithmetic modules we presented.

We implement program that approximates Euler's constant. Euler's

```
crn = {
  conc[one,1], conc[n,n0],
  step[{
    Add[z,one,znext],
    Mul[znext,znext,zpow],
    Cmp[zpow,n]
  }],
  step[{
    IfLT[{Ld[znext,z]}],
    IfGE[{Ld[z,out]}]
  }]
};
```

(a) *CRN++* code



(b) Simulation results for $n0 = 10$. Values of z (green), $zpow$ (orange), and out (red) are shown.

Figure 4.5: Integer square root in *CRN++*

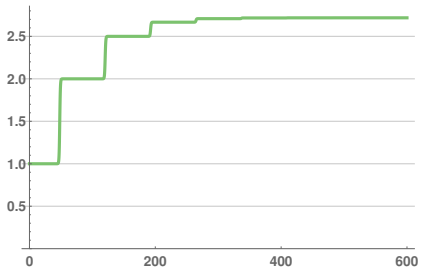
constant can be computed using the following infinite series:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{1} + \frac{1}{1} + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} + ...$$

We map this program into *CRN++* code, as shown in Fig 4.6. Variable e contains current approximation of the constant, while *element* stores the current element of the series. In the first step (Lines 7-11), the *element* is divded by the *divisor*, *divisor* incremented for the next iteration, and e incremented by the current element of the series. In the second step (Lines 12-16), the temporary variables $elementNext$, $eNext$, and $divisorNext$, are restored into the original ones. Precision achieved at the end of simulation is up to 5 decimal digits, we get result 2.71828.

```
crn = {
  conc[e, 1],
  conc[element, 1],
  conc[divisor, 1],
  conc[one, 1],
  conc[divisorMultiplier, 1],
  step[{
    Dvd[element, divisor, elementNext],
    Add[divisor, one, divisorNext],
    Add[e, elementNext, eNext]
  }],
  step[{
    Ld[elementNext, element],
    Ld[divisorNext, divisor],
    Ld[eNext, e]
  }]
};
```

(a) *CRN++* code



(b) Simulation results; value of e is shown (green line).

Figure 4.6: Approximating Euler's constant in *CRN++*

4.1.6 Approximating π

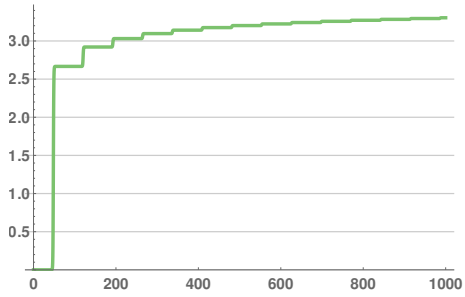
We approximate the π constant via *CRN++* program. We rely on the following infinite series to do so:

$$\pi = \frac{4}{1} - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \frac{4}{9} - \frac{4}{11} + ...$$

Fig 4.7 shows both code and simulation. In the first step (Lines 6-13), 4 is divided by the current divisor $divisor1$ and stored into the $factor1$, also 4 is divided by the $divisor2 := divisor1 + 2$ and stored into the $factor2$, $factor1$ and $factor2$ are subtracted and added to pi, at the same time $divisor1$ and $divisor2$ are increased by 2 for the next iteration. In the second step (Lines 14-18), the temporary variables $divisor1Next$, $divisor2Next$, and pi are restored to the original variables. Value of pi at the end of simulation is 3.1417. Note

```
crn={
  conc[four, 4],
  conc[divisor1, 1],
  conc[divisor2, 3],
  conc[pi, 0],
  step[{
    Dvd[four, divisor1, factor1],
    Add[divisor1, four, divisor1Next],
    Dvd[four, divisor2, factor2],
    Add[divisor2, four, divisor2Next],
    Sub[factor1, factor2, factor],
    Add[pi, factor, piNext]
  }],
  step[{
    Ld[divisor1Next, divisor1],
    Ld[divisor2Next, divisor2],
    Ld[piNext, pi]
  }]
};
```

(a) *CRN++* code



(b) Simulation results; value of pi is shown (green line).

Figure 4.7: Approximating number π in *CRN++*

that error builds up, if we increase simulation time π converges to value that is in $\epsilon = 0.2$ range of the correct result. This is unlike the approximation of the Euler's constant; error evaluation shows that the reason is due to using the subtraction (of close values) to approximate the π, and subtraction is the most error-prone operation out of all arithmetic modules we present (see Section 4.2).

4.2 Error Characterization

In this section we evaluate the error of the basic arithmetic modules (Section 4.2.1), and present the idea of redesigning CRNs to reduce the er-

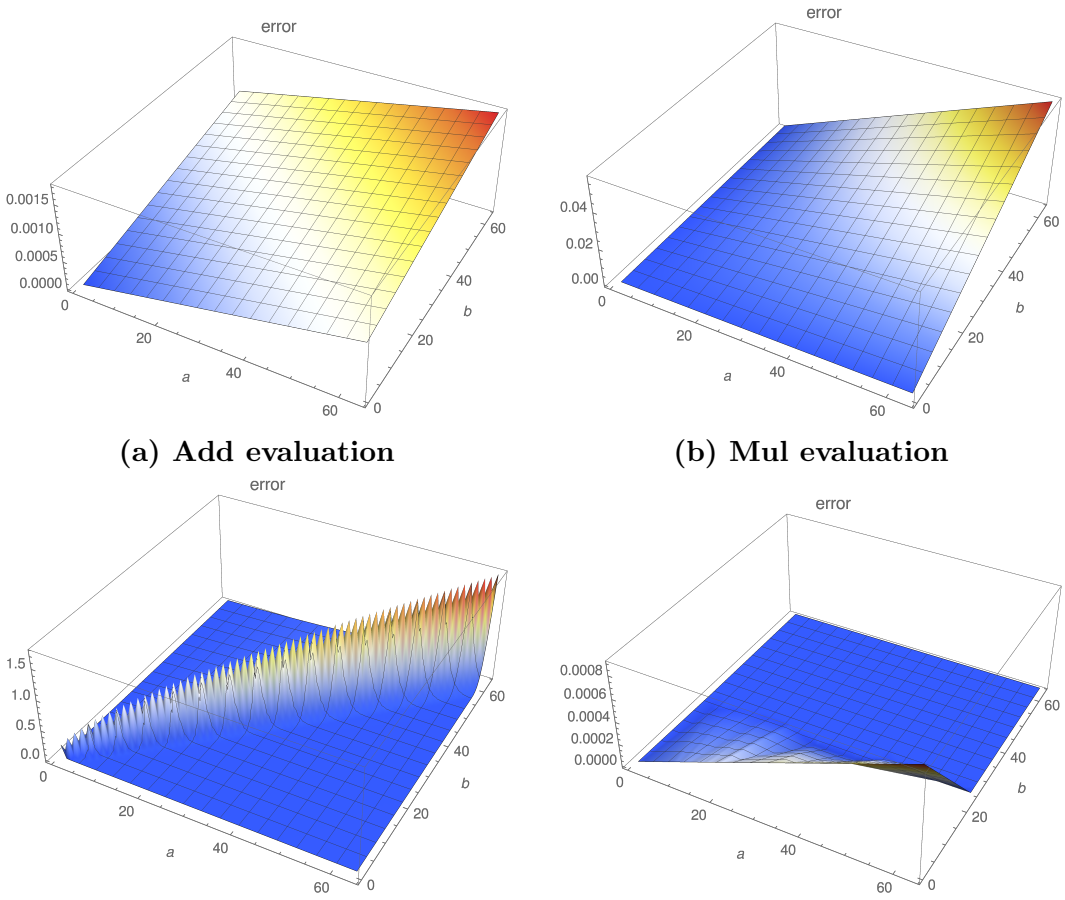
ror (Section 4.2.2).

4.2.1 Error of Arithmetic Modules

Using our error evaluation mechanisms (Section 3.4) we analyze the error of the modules. We evaluate each module separately, on different inputs, to characterize its behavior. Figure 4.8 shows the error evaluation results, x and y axis reflect values of the first (a) and second (b) operand, respectively, while z axis shows the error. The plots provide useful knowledge: (a) The *Mul* module error depends on the value being computed, it increases as the value being computed increases, and does not depend on the order of arguments—preserves commutativity property; (b) The *Add* module follows the same pattern as *Mul*, but has a lower absolute error; (c) The *Sub* exhibits the maximum error when inputs are close to each other, and in general, has higher error rate than other arithmetic modules. This knowledge is useful when designing *CRN++* programs, we realize that the particularly error-prone operation is subtraction of the arguments close to each other; this is indeed the reason why error in example approximating π constant (4.1.6) is higher than in the one approximating Euler's number (4.1.5). Having this in mind, a user can decide to rewrite programs in more optimal way; for example, the subtraction of close operands can be done in alternative, less error-prone way (Figure 4.9b). We plan to add runtime assertions to *CRN++* programs that alert for possible issues in the program, for example, when values being subtracted are closer than ϵ to each other.

4.2.2 Reducing the error through program refactoring

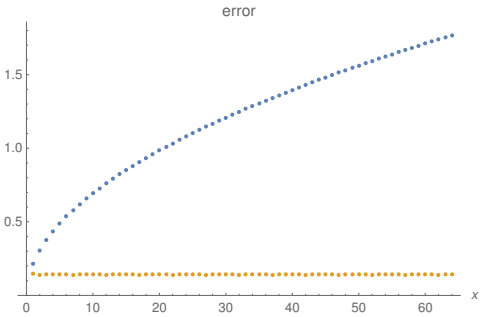
The *Sub* has a high error when operands are close to each other; but there are alternative ways to subtract. Figure 4.9b shows the alternative code for performing subtraction. Value of b is subtracted from a, by iteratively subtracting 1 from both a and b, until b reaches 0.



(a) Add evaluation **(b) Mul evaluation**

(c) Sub evaluation **(d) Dvd evaluation**

Figure 4.8: Error evaluation of arithmetic modules.
Axis a and b show the values of the first and second operand, respectively; z axis show the value of the error (difference between the correct and actual value of the operation).



(a) Comparing error of Sub module (blue lines) and alternative way to subtract (orange lines). X-axis show the value of both minuend and subtrahend.

```
crn = {
  conc[a, a0], conc[b, b0],
  conc[one, 1], conc[zero, 0],
  step[{
    Cmp[b, zero]
  }],
  step[{
    IfGE[{
      Sub[a, one, anext],
      Sub[b, one, bnext]
    }]
  }],
  step[{
    IfGE[{
      Ld[anext, a],
      Ld[bnext, b]
    }]
  }]
}
```

(b) Alternative way to subtract

Figure 4.9: Error comparison of *Sub* and an alternative way for subtraction

Error evaluation is shown (Figure a) for the cases when the operands are equal (minuend and subtrahend same), since *Sub* exhibits the highest error in that case.

Chapter 5

Related Work

Computational power of chemical reaction networks. Previous research demonstrated techniques of achieving complex behaviors in chemistry, such as: computing algebraic functions [3], polynomials [21], implementing logic gates [20]. Moreover, the Turing completeness of chemistry has been proven, using the strategy of implementing polynomial ODEs (which have been previously shown to be Turing universal) in mass-action chemical kinetics [9]. Even though Turing complete, this translation to chemistry can result in infeasibly complex chemical reaction networks, which motivates other, more direct methods.

Modular Reactions. Adding even a single reaction to a CRN can completely change its dynamics, which makes the design process challenging. The idea of 'composable' reactions seeks a set of reactions that can be composed in a well-defined manner to implement more complex behaviors. Buisman et al. [3] compute algebraic expressions by designing the core modules that implement basic arithmetic operations, which can be further composed to achieve more complex tasks. Our goal is to make modular designs, and we follow some of the proposed design principles for achieving the goal, such as

input-preserving CRNs.

Synchronous computation. Previous work utilized synchronous logic to achieve complex tasks. Soloveichik et al. implement state machines in chemistry by relying on clock species [25]. We use the same technique, where we add clock species acting catalytically to order reactions. Jiang et al.[15], also relying on clock species, design a model of memory in chemistry to support sequential computation, demonstrating their technique on examples of a binary counter and a fast Fourier transform (FFT). Previous work shows the promise of programming synchronous logic in reactions, which we advance by providing an explicit programming language and framework for designing and testing wide-range of programs.

Asynchronous computation. Recall, an absence indicator is a species that is present in high concentration when a target species is present in low concentration. Absence indicators can be used to drive a reaction when a particular reaction has finished, providing a method for executing modules in desired order. Huang et al. [14] use absence indicators to implement algorithms such as integer division and GCD. Their method requires two reaction rates, 'fast' and 'slow', where the fast rate needs to be orders (2-3) of magnitude larger to ensure the proper function of the system. Since, in practice, biochemical systems allow for a restricted range of reaction rates, requiring a large spectrum of rates slows down the computation when the computation speed is dictated by the slow rates. In contrast, we allow all reactions to take the same (or comparable) rate constants. While the goal of our work is not

to compare asynchronous and synchronous computation, we mention insights and intuition of their differences, which we gained through empirical studies. First, absence indicators are not robust, and typically require fine tuning to get the system right. Second, error detection is easier with synchronous logic. Since all operations follow the clock signal, there is a direct mapping from a time moment to a command that is executing, which provides a way to check correctness of a system at any point of time. Finally, we provide a framework for implementing molecular programs which is easily extensible, and can be used to compare synchronous and asynchronous logic. We include support for absence indicators through a *IfAbsent* construct, thus allowing easy comparison of the two paradigms.

Chapter 6

Discussion and conclusions

There are multiple ways in which we can further improve *CRN++*. Note that currently every high-level module is mapped to exactly one CRN implementing the operation. Letting the tool decide which implementation to use in different contexts could boost the performance. For example, the described modules have a useful property of preserving inputs, but that property might not be needed in every case. If the input preserving property is redundant, *CRN++* could choose to use the more optimized version (for example the more compact subtraction CRN discussed above). Also, we can improve the programming experience by allowing the same species as both input and output of a module, and do the background work to allocate temporary variables.

An important direction for future research concerns reducing the error in our construction, and how it builds up over time. We noticed that different algorithms, even computing the same function, accumulate varying levels of error. For example, as seen in 4.2.1, the error of the Sub module increases with the magnitude of the operands, and also increases the closer they are. However, we also found an alternative way to subtract, that keeps the error constant and independent of the operands (see Figure 4.9b) at the cost of

slower run-time. Our error analysis shows that for most examples we tried, but not all, error builds up over the course of the computation. For the *CRN++* programs where the error builds up in this way, there is some maximum input complexity beyond which the error overwhelms the output. Can all *CRN++* programs be refactored (preferably automatically) to bound the cumulative error of every module such that it does not build up over time? Note that if this were possible, we would obtain another, more efficient, way to achieve Turing universality.

To the best of our knowledge we are the first to provide an imperative programming language which compiles to chemical reaction networks. Moreover, we build tools that can help users get a better understanding of CRNs and improve their design. Although, absolutely correct computation is not achieved, we provide tools that help understand why error occurs and improve the design of CRNs. We release our toolkit as open source, to encourage new research and improvement of the *CRN++*, with the hope of advancing the engineering of information processing molecular systems.

Bibliography

[1] Dana Angluin, James Aspnes, and David Eisenstat. A simple population protocol for fast robust approximate majority. *Distributed Computing*, 21(2):87–102, 2008.

[2] Olivier Bournez, Daniel S Graça, and Amaury Pouly. Polynomial time corresponds to solutions of polynomial ordinary differential equations of polynomial length: the general purpose analog computer and computable analysis are two efficiently equivalent models of computations. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 55. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[3] H. J. Buisman, Huub M. M. ten Eikelder, Peter A. J. Hilbers, and Anthony M. L. Liekens. Computing algebraic functions with biochemical reaction networks. *Artificial Life*, pages 5–19, 2009.

[4] Luca Cardelli, Milan Češka, Martin Fränzle, Marta Kwiatkowska, Luca Laurenti, Nicola Paoletti, and Max Whitby. *Syntax-Guided Optimal Synthesis for Chemical Reaction Networks*, pages 375–395. 2017.

[5] Luca Cardelli and Attila Csikász-Nagy. The cell cycle switch computes approximate majority. *Scientific reports*, 2:656, 2012.

[6] Yuan-Jyue Chen, Neil Dalchau, Niranjan Srinivas, Andrew Phillips, Luca Cardelli, David Soloveichik, and Georg Seelig. Programmable chemical controllers made from DNA. *Nature nanotechnology*, 8(10):755, 2013.

[7] CRNSimulator Mathematica package. `http://users.ece.utexas.edu/~soloveichik/crnsimulator.html`.

[8] Peter Dittrich, Jens Ziegler, and Wolfgang Banzhaf. Artificial chemistries – a review. *Artificial life*, 7(3):225–275, 2001.

[9] François Fages, Guillaume Le Guludec, Olivier Bournez, and Amaury Pouly. Strong Turing completeness of continuous chemical reaction networks and compilation of mixed analog-digital programs. In *International Conference on Computational Methods in Systems Biology*, pages 108–127, 2017.

[10] Brian Fett, Jehoshua Bruck, and Marc D Riedel. Synthesizing stochasticity in biochemical systems. In *Proceedings of the 44th annual Design Automation Conference*, pages 640–645, 2007.

[11] Brian Fett and Marc D Riedel. Module locking in biochemical synthesis. In *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, pages 758–764, 2008.

[12] Jinting Gao, Yaqing Liu, Xiaodong Lin, Jiankang Deng, Jinjin Yin, and Shuo Wang. Implementation of cascade logic gates and majority logic

gate on a simple and universal molecular platform. *Scientific reports*, 7(1):14014, 2017.

[13] Benjamin Groves, Yuan-Jyue Chen, Chiara Zurla, Sergii Pochekailov, Jonathan L Kirschman, Philip J Santangelo, and Georg Seelig. Computing in mammalian cells with nucleic acid strand exchange. *Nature nanotechnology*, 11(3):287, 2016.

[14] De-An Huang, Jie-Hong R. Jiang, Ruei-Yang Huang, and Chi-Yun Cheng. Compiling program control flows into biochemical reactions. In *Proceedings of the International Conference on Computer-Aided Design*, pages 361–368, 2012.

[15] H. Jiang, M. Riedel, and K. Parhi. Synchronous sequential computation with molecular reactions. In *2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 836–841, 2011.

[16] H. Jiang, M. Riedel, and K. Parhi. Synchronous sequential computation with molecular reactions. In *2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 836–841, 2011.

[17] Ahmad S Khalil and James J Collins. Synthetic biology: applications come of age. *Nature Reviews Genetics*, 11(5):367, 2010.

[18] Michael Lachmann and Guy Sella. The computationally complete ant colony: Global coordination in a system with no hierarchy. In *European Conference on Artificial Life*, pages 784–800. Springer, 1995.

[19] Alex Lake, Stephen Shang, and Dmitry M Kolpashchikov. Molecular logic gates connected through DNA four-way junctions. *Angewandte Chemie International Edition*, 49(26):4459–4462, 2010.

[20] Marcelo O Magnasco. Chemical kinetics is Turing universal. *Physical Review Letters*, 78(6):1190, 1997.

[21] Sayed Ahmad Salehi, Keshab K. Parhi, and Marc D. Riedel. Chemical reaction networks for computing polynomials. *ACS Synthetic Biology*, 6(1):76–83, 2017.

[22] Phillip Senum and Marc Riedel. Rate-independent constructs for chemical computation. *PloS one*, 2011.

[23] Adam Shea, Brian Fett, Marc D Riedel, and Keshab Parhi. Writing and compiling code into biochemistry. In *Biocomputing 2010*, pages 456–464. 2010.

[24] David Soloveichik, Matthew Cook, Erik Winfree, and Jehoshua Bruck. Computation with finite stochastic chemical reaction networks. *natural computing*, 7(4):615–633, 2008.

[25] David Soloveichik, Georg Seelig, and Erik Winfree. DNA as a universal substrate for chemical kinetics. *Proceedings of the National Academy of Sciences*, 107(12):5393–5398, 2010.