PLOS ONE

RESEARCH ARTICLE

Obstructions to Sampling Qualitative Properties

Arne C. Reimers*

Life Sciences Group, Centrum Wiskunde & Informatica, Amsterdam, Netherlands

* arne.c.reimers@gmail.com

Citation: Reimers AC (2015) Obstructions to Sampling Qualitative Properties. PLoS ONE 10(8): e0135636. doi:10.1371/journal.pone.0135636

Editor: Peter Csermely, Semmelweis University, HUNGARY

Received: March 10, 2015

Accepted: July 23, 2015

Published: August 19, 2015

Data Availability Statement: The paper only uses already published data. The sourcecode used for the analysis is part of supporting information.

Funding: This work was supported by Berlin Mathematical School (PhD-stipend) (www.math-berlin.de) and European Research Consortium for Informatics and Mathematics ("Alain Bensoussan Fellowship Programme") (https://fellowship.ercim.eu/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Background

Sampling methods have proven to be a very efficient and intuitive method to understand properties of complicated spaces that cannot easily be computed using deterministic methods. Therefore, sampling methods became a popular tool in the applied sciences.

Results

Here, we show that sampling methods are not an appropriate tool to analyze qualitative properties of complicated spaces unless **RP** = **NP**. We illustrate these results on the example of the thermodynamically feasible flux space of genome-scale metabolic networks and show that with artificial centering hit and run (ACHR) not all reactions that can have variable flux rates are sampled with variables flux rates. In particular a uniform sample of the flux space would not sample the flux variabilities completely.

Conclusion

We conclude that unless theoretical convergence results exist, qualitative results obtained from sampling methods should be considered with caution and if possible double checked using a deterministic method.

Introduction

Given a space $S \subseteq \mathbb{R}^n$, we are interested in computing a set of *sample points* $s_1, \ldots, s_k \in S$ that represent the space S and its properties. Thus, by randomly generating sample points this offers one approach to overcome the curse of dimensionality. For example, if S is a polyhedron, we can compute nearly uniformly distributed sampling points of S in polynomial time [1] and from this approximate the volume of the polyhedron [2, 3]. In contrast no deterministic polynomial-time algorithm can compute the volume of convex sets with less than exponential relative error in n [4, 5].

Thus, *sampling methods* are nowadays used in many application areas, for example in the analysis of flux spaces in genome-scale metabolic networks. A genome-scale metabolic network models the chemical reactions possible in an organism. Using the assumption that no

internal substance can be over- or under-produced a flow problem is obtained. This then leads to a polyhedron of feasible flows (also called fluxes) through the network (called flux space):

$$F := \{v \in \mathbb{R}^n : Sv = 0, \ell \leq v \leq u\}$$

Here, $S \in \mathbb{R}^{m \times n}$ is called the stoichiometric matrix. It encodes for each reaction $r \in \mathcal{R} = \{1, \ldots, n\}$ which and how much of the metabolites $\mathcal{M} = \{1, \ldots, m\}$ are consumed resp. produced. ℓ, u are bounds on the reaction rates. Due to the size of these networks, deterministic methods to enumerate extreme points and thus a representative set of feasible flows [6, 7] are unpractical.

Therefore, tools and methods have been developed to sample the whole flux space [8–11] or only the extreme points [12] and then used to derive biological insights [13–16]. Typical properties that are analyzed are correlations between fluxes through different reactions [17–20], or the distribution of flux rates through a given reaction [21]

While points in polyhedra can be sampled efficiently (in theory), often additional constraints are added to make the results more biological reasonable. However, this often makes the space of feasible solutions non-convex and associated decision problems **NP**-hard. This is for example the case with "looplaw"-thermodynamic constraints [21, 22], here sloppily represented using the phrase "v thermo. feasible":

$$T := \{v \in F : v \text{ thermo. feasible}\}$$

To test how well sampling works, we consider the scenario that we want to decide if the flux space has a given property. In Practical Obstructions to Sampling we consider the problem of determining if positive resp. negative flux through a reaction is possible. There, we show that with *artificial centering hit and run* (ACHR) [8, 23] we cannot determine this property correctly for several reactions in genome-scale metabolic networks. In particular, the artifacts are not only observed in non-convex flux spaces, but also for polyhedral flux spaces.

In Theoretical Obstructions to Sampling we show that for the non-convex flux space T this problem is not specific to ACHR, but more fundamental. Therefore, we generalize the problem to decide if a space S has a given property (formulated as a decision problem Prob) by using a sampling algorithm. We define the concept of *non-trivial polynomial time sampling algorithm* and show how it can be used to solve decision problems in randomized polynomial time. We show that if the decision property is NP-hard, then there exists no polynomial time sampling algorithm that samples S in a non-trivial way w.r.t. to the property formulated by Prob unless **NP** = **RP**, where **RP** is the class of problems that can be solved in randomized polynomial time [24].

Practical Obstructions to Sampling

Thermodynamic constraints are an additional source of constraints that have been used in the analysis of metabolic networks [25–34] and were also used in sampling methods [21, 35, 36]. However, most thermodynamic constraints are computationally difficult, since they are non-convex. Here we use so-called "looplaw"-thermodynamic constraints.

For the toy network shown in Fig 1 "looplaw"-thermodynamic constraints imply that positive flux through r_1 and r_2 at the same time is not possible, because they form a stoichiometrically balanced internal cycle. Similarly, positive flux through r_1, r_3, and r_4 at the same time is also not possible. This implies that the flux space looks as shown in Fig 2. Clearly, this flux space is not convex. Furthermore, we observe that if we sample the flux space uniformly, we would not sample any flux with positive flux through r_1, since the flux space with positive flux through r_1 is 1-dimensional while the rest of the flux space is 2-dimensional.

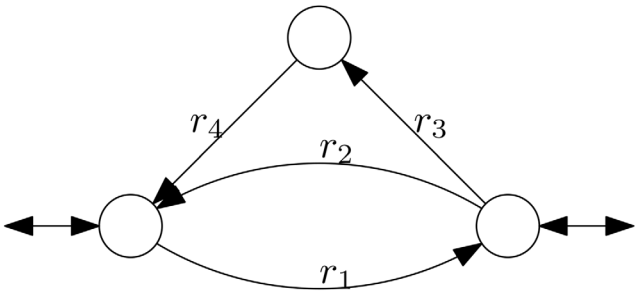

Fig 1. Toy network. Internal reactions r_1, r_2, r_3, r_4 are irreversible. By thermodynamics, it is not possible to have non-zero flux through r_1 and also to have a non-zero flux through one of r_2, r_3 or r_4 at the same time.

doi:10.1371/journal.pone.0135636.g001

For "looplaw"-thermodynamic constraints we showed that deciding if a reaction can carry positive flux is **NP**-hard [37]. Thus, according to our theoretical results in Theoretical Obstructions to Sampling, it should be harder to sample flux spaces with thermodynamic constraints than without. Therefore, we test how well we can predict the reactions with variable flux rates by sampling fluxes through the network. Finding all reactions with variable flux rate is called *flux variability analysis* (FVA). If it is applied without thermodynamic constraints, it can be solved efficiently using linear programming [38, 39]. With thermodynamic constraints, we can typically solve it in practice using mixed integer linear programming techniques (MILP) [21, 37]. For a good sampling algorithm we expect that for every reaction that can have positive flux (as determined by FVA) we also get samples whith positive flux through the reaction. Therefore, we count for how many of the reactions we fail this goal, i.e. where we do not even obtain a single sample with a positive flux through the reaction.

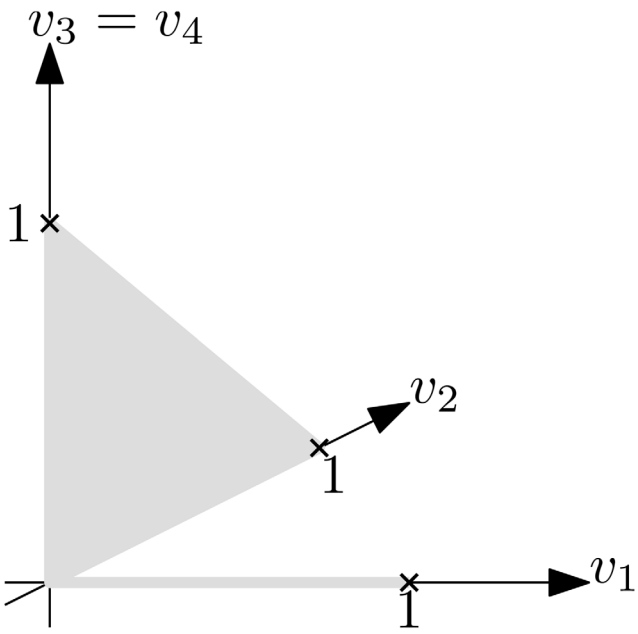

Fig 2. Flux space of toy network. The gray area denotes the flux space. In this example it was assumed that input/output flux values are constrained to at most 1. It can be seen that flux v_1 through r_1 is exclusive to fluxes v_2 and v_3 through r_2 and r_3 respectively. Since fluxes through r_2 can be combined with fluxes through r_3, the flux space with v_2, $v_3 > 0$ is two-dimensional, while the flux space with $v_1 > 0$ is only one-dimensional. Hence, a uniform sample of the flux space would almost surely have zero flux through r_1.

doi:10.1371/journal.pone.0135636.g002

Of particular interest are the reactions $\mathcal{R}^C$ of reactions contained in internal cycles, because they are affected by thermodynamic constraints. The reader is referred to [37] for more details and a precise definition. The reactions not contained in internal cycles $\mathcal{R}^{NC}$ on the other hand should not be affected by "looplaw"-thermodynamic constraints [37]

Method

To verify the impact of the theoretical results, we implemented the following computational experiment to analyze the difference between sampling with thermodynamic constraints and without thermodynamic constraints. For a given metabolic network with flux space P (with or without thermodynamic constraints) we do the following:

1. Sample n points in the flux space P.
2. Run flux variability analysis (FVA) on P and define:
   - $\mathcal{R}_+ :=$ reactions that can have positive flux,
   - $\mathcal{R}_- :=$ reactions that can have negative flux.
3. From this we define the following 4 reaction classes:
   - $\mathcal{R}_+^C := \mathcal{R}^C \cap \mathcal{R}_+$
   - $\mathcal{R}_+^{NC} := \mathcal{R}^{NC} \cap \mathcal{R}_+$
   - $\mathcal{R}_-^C := \mathcal{R}^C \cap \mathcal{R}_-$
   - $\mathcal{R}_-^{NC} := \mathcal{R}^{NC} \cap \mathcal{R}_-$
4. For each reaction class $A \subseteq \mathcal{R}_+$, we count the number of reactions for which we never sampled positive flux n_A^P and then compute the ratio $r_A^P := \frac{n_A^P}{|A|}$.
5. For each reaction class $A \subseteq \mathcal{R}_-$, we count the number of reactions for which we never sampled negative flux n_A^P and then compute the ratio $r_A^P := \frac{n_A^P}{|A|}$.

We do this for the steady-state flux space F (without thermodynamic constraints) and for the thermodynamically constrained flux space T. Since positive lower bounds and negative upper bounds for reactions in internal cycles make it already **NP**-hard to find a thermodynamically feasible flux distribution, we set all positive lower bounds and all negative upper bounds to 0.

For the sampling method we chose to use the ACHR method implemented in the COBRA toolbox [40], since it is one of the most established tools for sampling flux spaces. They also offer a flag to activate thermodynamic constraints. Unfortunately this flag has no effect in the current version (2.0.5). Hence, we implemented a simple post-processing step to turn thermodynamically infeasible fluxes into thermodynamically feasible fluxes by deleting internal cycles [37]. To check that our results are not an artifact of our post-processing step, we also implemented the post-processing method suggested by Schellenberger et al. [21], where for each sample point a thermodynamically feasible flux vector is computed that minimizes the L^1-norm distance to the sample point. We remark that this method solves an MILP in the post-processing step and hence, cannot be considered a polynomial-time sampling method.

The sampling method was run with default parameters, except that the number of points per file is half the number of output points. This means that for each sample point ACHR performed at least 200 steps and potentially biased samples from the beginning were dropped. We choose 10000 output points, since this allowed to run the analysis in a couple of hours. In

contrast, the variability of reactions (with and without thermodynamic constraints) can be computed deterministically using the FVA-method in [37] in a few minutes.

We selected a set of genome-scale metabolic networks based from the BiGG-database [41] as a test-set, since these networks are well curated and well established test-networks. We did not select *Human Recon 1*., since we were not able to run thermodynamically constrained flux variability analysis on it. Instead, we also added the more recent *E. coli* iJO1366 network [42]. Also, we did not select the *M. barkeri* network because the sampling algorithm from the COBRA toolbox crashed. The matlab scripts for the computations can be found in S1 code.

Results

The computed results for 10000 samples using the cycle deletion method [37] can be seen in Fig 3. The L^1-minimization method [21] was not practically applicable for most of the test networks, because it took more than 10 minutes to compute the closest thermodynamically feasible flux vector for even the first sample point. Only for *H. pylori* iIT341 and *S. cerevisiae* iND750 this was a feasible approach. For these networks, however, already the simple cycle deletion method already managed to sample non-zero fluxes for all reactions in internal cycles that can have non-zero flux. Hence, these results are not separately shown.

We observe that the ratio of reactions where no positive/negative fluxes were sampled is larger for the case with thermodynamic constraints than without. Also, as expected, the results for $\mathcal{R}^{NC}$ are the same for *F* and *T*.

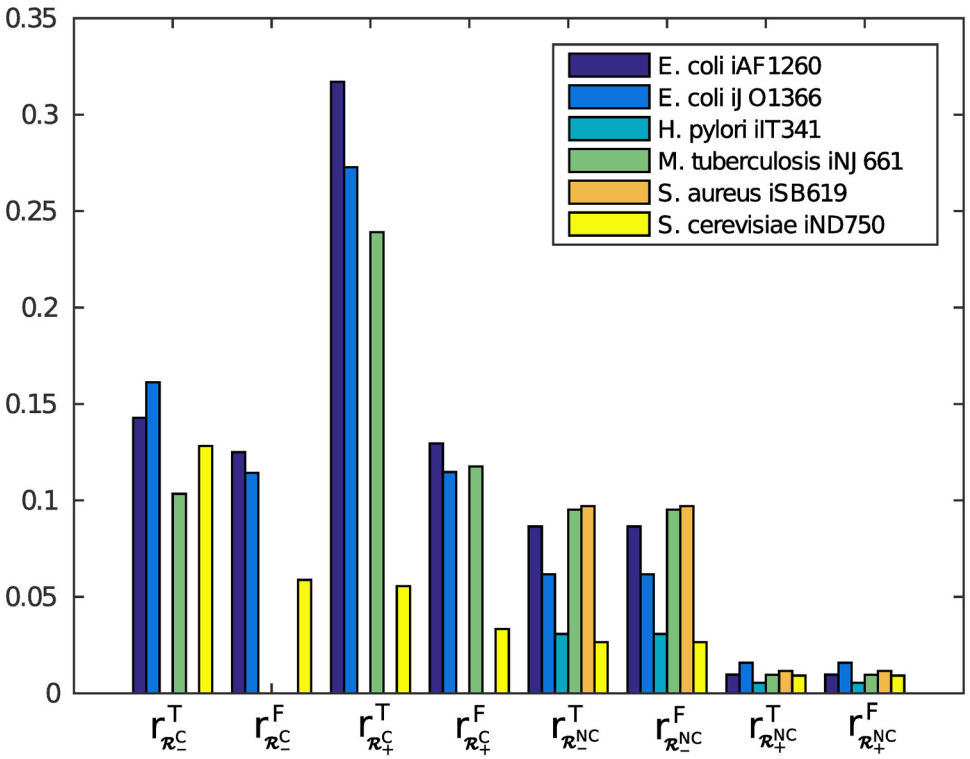

Fig 3. Sampling results with 10000 sample points. The *y*-axis shows the ratio for how many reactions that can have positive/negative flux the sampling method did not sample at least one such flux vector.

doi:10.1371/journal.pone.0135636.g003

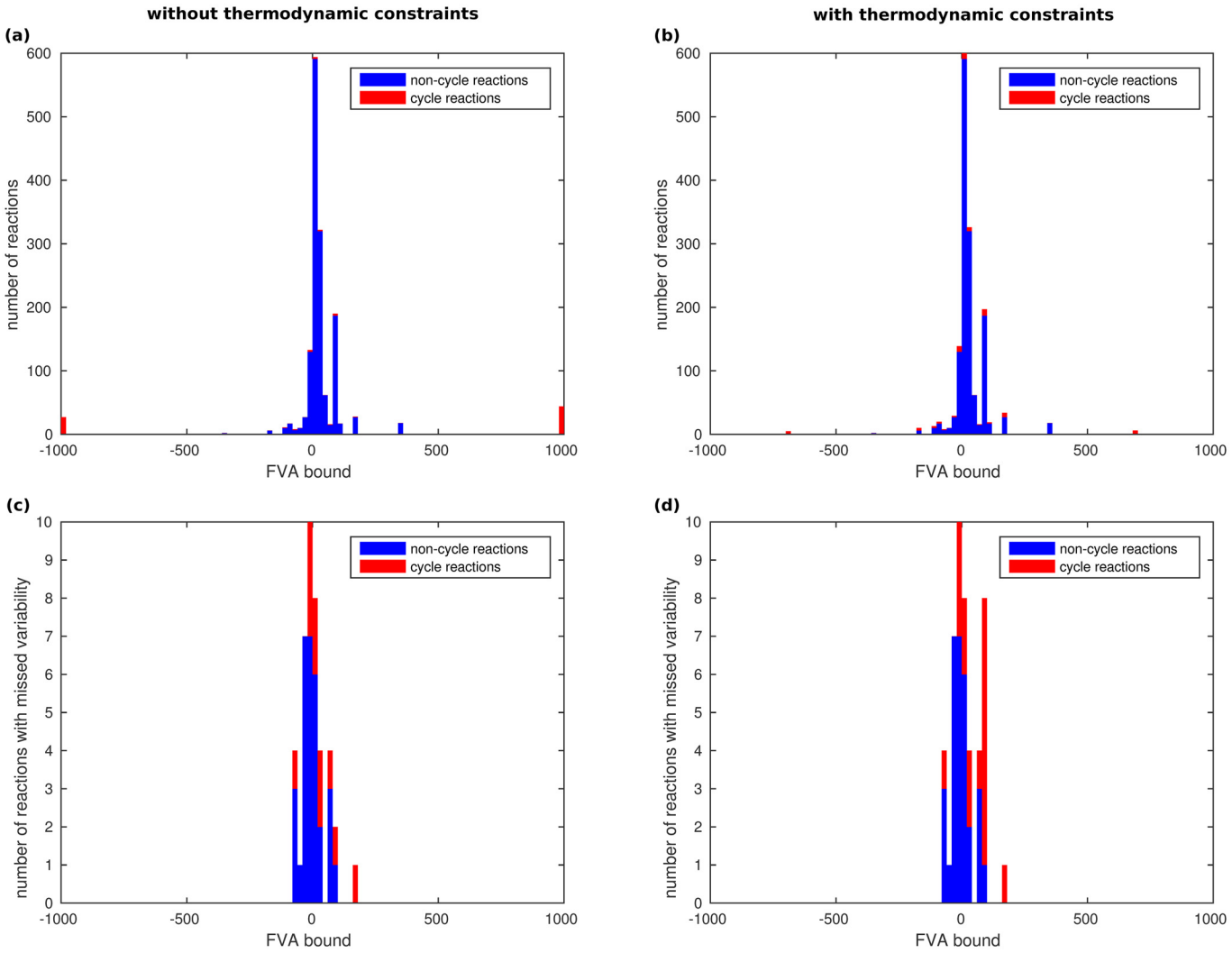

Fig 4. Distribution of *E. coli* iAF1260 flux variability that is missed by sampling. Histograms (a) and (c) are for the flux space without thermodynamic constraints, histograms (b) and (d) are with thermodynamic constraints. The bins in histograms (a) and (b) count the number of reactions with the respective lower and upper bounds computed by FVA. Bounds equal to 0 are not counted. Histograms (c) and (d) show, for the lower and upper bounds shown in (a) and (b), the number of reactions for which no negative resp. positive flux has been sampled. For all histograms, bin sizes of length 20 were chosen. We remark that zero flux is always possible. Therefore, the flux bounds directly relate to the possible flux range.

doi:10.1371/journal.pone.0135636.g004

However, we are surprised to find that even without thermodynamic constraints, we miss about 5% of all possible reaction directions. This shows us that ACHR might not be a very good approach for sampling steady-state flux spaces of genome-scale metabolic networks, as has also been observed in [43].

One potential cause for the bad performance of ACHR is that the flux space of metabolic networks is badly conditioned, i.e., that there exist reactions with very low variability and reactions with high variability. While this is indeed the case (see Fig 4a and 4b), it does not appear to be the primary cause, since not only reactions with very low variability are missed (see Fig 4c and 4d). From Fig 4 we also see that there exists an assymmetry between lower flux bounds and upper flux bounds, which could be a reason for the different results for positive and negative directions in Fig 3.

Theoretical Obstructions to Sampling

Let PROB:$\mathcal{J} \to \{0, 1\}$ be an **NP**-hard decision problem on a set $\mathcal{J}$ of inputs (commonly we use the set of words over the alphabet {0, 1} as input, i.e., $\mathcal{J} = \{0, 1\}^*$ and the length of an input is just the length of the word). To solve PROB by sampling, we require that the structure of the sampling space represents PROB in a certain way. This we encode using a function $\mathcal{X}$ that maps every input $I \in \mathcal{J}$ into a subset of $\mathbb{R}^n$, i.e. an element of the powerset of $\mathbb{R}^n$ ($\mathcal{P}(\mathbb{R}^n)$). This space $\mathcal{X}(I)$ will be the space from which we will draw samples. Note that we will make no assumptions on the size of n compared to the input size $|I|$. Additionally, we use a test-function f that tests whether a point of the sample-space $x \in \mathcal{X}(I)$ has a non-trivial property, i.e. $f(I, x) > 0$.

Definition 1 (Sampling space) Given a decision problem PROB:$\mathcal{J} \to \{0, 1\}$, we call $(\mathcal{X}, f)$ a *sampling space* for PROB if $\mathcal{X}: \mathcal{J} \to \mathcal{P}(\mathbb{R}^n)$ and $f: \mathcal{J} \times \mathbb{R}^n \to \mathbb{R}$ satisfy:

- $f(I, x)$ is continuous in x for all $x \in \mathbb{R}^n$.
- $f(I, x)$ can be computed in time polynomial in the encoding length of I and x. For $x \in \mathbb{R}$ without a finite encoding length, we assume that $f(I, x)$ is well defined, but its computation does not terminate.
- It holds for all $I \in \mathcal{J}$ that

$$\mathrm{PROB}(I) = \begin{cases} 1 & \exists x \in \mathcal{X}(I) : f(I, x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

For example, let us assume that we want to know whether a given reaction $r \in \mathcal{R}$ can have positive flux in the flux space F of a given metabolic network. Thus, a problem instance I encodes a metabolic network $\mathcal{N}(I)$ and a target reaction $r(I)$. It follows that $(\mathcal{X}, f)$ with $\mathcal{X}(I) \coloneqq F(I)$ and $f(I, x) \coloneqq \mathrm{pr}_{r(I)}(x)$ for all $I \in \mathcal{J}$ is a sampling space, where $F(I)$ is the flux space of the metabolic network $\mathcal{N}(I)$ and $\mathrm{pr}_r(x)$ denotes the flux through reaction r in the flux vector x.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, i.e., a sample space Ω, events $\mathcal{F}$, and probability function P. It will serve us as the space from which we draw the seeds for the sampling algorithm. Here, we assume that the sampling method is given as a function $\mathcal{S}: \mathcal{J} \times \mathbb{N} \times \Omega \to \mathbb{R}^n$, i.e., for every time point we get a sample. With this formalism we want to capture the behavior of random-walk sampling methods that do a random walk through $\mathcal{X}(I)$ and can be run for arbitrarily long times to improve the sampling result. Classical sampling algorithms can also be captured by this formalism by iteratively running the sampling method and computing a consensus value. If the sampling algorithm did not produce a result for an (early) time point, it could simply return a default value. Since we will only consider asymptotic behavior, this will not be of any importance.

Definition 2 (Feasible Sampling Algorithm) $\mathcal{S}: \mathcal{J} \times \mathbb{N} \times \Omega \to \mathbb{R}^n$ is a *feasible sampling algorithm*, if there exists a polynomial $p: \mathbb{N} \to \mathbb{R}$ such that

$$\mathcal{S}(I, k, \omega) \in \mathcal{X}(I) \quad \forall k \geq p(|I|), I \in \mathcal{J}, \omega \in \Omega$$

Definition 3 (Polynomial Time Sampling Algorithm) $\mathcal{S}: \mathcal{J} \times \mathbb{N} \times \Omega \to \mathbb{R}^n$ is a *polynomial time sampling algorithm* if there exists a polynomial $q: \mathbb{N} \times \mathbb{R}^+ \to \mathbb{R}^+$ and for every $I \in \mathcal{J}$ a random variable $X: \Omega \to \mathbb{R}^n$ such that

- $\mathcal{S}(I, k, \omega)$ for $I \in \mathcal{J}$ and $\omega \in \Omega$ can be computed in time $O(k)$,
- $\mathcal{S}(I, k, \cdot) \to X$ in distribution for $k \to \infty$, and

- $\mathcal{S}(I, k, \cdot)$ converges to X in polynomial time, i.e., for every closed set $A \subseteq \mathbb{R}^n$ holds

$$|P(\mathcal{S}(I, k, \cdot) \in A) - P(X \in A)| < \varepsilon$$

for all $k > q(|I|, \varepsilon^{-1})$.

Assume there exists such a sampling method $\mathcal{S} : \mathcal{J} \times \mathbb{N} \to \mathbb{R}^n$ that samples the feasibility space $\mathcal{X}(I)$ of the **NP**-hard optimization problem Prob for each given instance $I \in \mathcal{J}$ in a non-trivial way, i.e., without losing any features (represented by f):

Definition 4 (Non-trivial Sampling Algorithm) $\mathcal{S} : \mathcal{J} \times \mathbb{N} \times \Omega \to \mathbb{R}^n$ is a *non-trivial sampling algorithm* w.r.t. $f : \mathcal{J} \times \mathbb{R}^n \to \mathbb{R}$ if for every $I \in \mathcal{J}$ there exists a random variable $X : \Omega \to \mathbb{R}^n$ such that

- $\mathcal{S}(I, k, \cdot) \to X$ in distribution for $k \to \infty$.
- If $\exists x \in \mathcal{X}(I)$ with $f(I, x) > 0$, then

$$P(f(I, X) \leq 0) = t < 1$$

with $\frac{1}{1-t} \leq p(|I|)$ for a polynomial p.

We can then use $\mathcal{S}$ to construct a probabilistic algorithm that will decide Prob. The probabilistic algorithm that we are going to construct will belong to the class **RP** (randomized polynomial time) [24].

Definition 5 (Complexity Class RP) A decision problem p is in **RP** if there exists a probabilistic algorithm that

- runs in polynomial time,
- if the answer to p is NO, it outputs NO, and
- if the answer to p is YES, it outputs YES with probability at least $\frac{1}{2}$.

Since **RP** = **NP** is an open problem in theoretical computer science, it is very unlikely that a given probabilistic polynomial-time sampling algorithm of the thermodynamically constrained flux space actually solves the **RP** = **NP** problem. Hence, it is much more likely that the sampling algorithm samples the feasible flux space incompletely.

Theorem 1 *Let* Prob: $\mathcal{J} \to \{0, 1\}$ *be an* **NP***-hard decision problem with sampling space* $(\mathcal{X}, f)$. *Unless* **RP** = **NP**, *there exists no feasible, non-trivial, polynomial time sampling algorithm* $\mathcal{S} : \mathcal{J} \times \mathbb{N} \times \Omega \to \mathbb{R}^n$.

Proof Assume there exists such a sampling algorithm. We construct an algorithm in **RP** for Prob.

For $I \in \mathcal{J}$ define $t(I) \coloneqq P(X \in A(I))$ for $A(I) \coloneqq \{x \in \mathbb{R}^n : f(I, x) \leq 0\}$. Since $f(I, \cdot)$ is continuous, it follows that $A(I)$ is closed and Borel-measurable. Hence, $t(I)$ is well defined.

By Def. 2 and Def. 3 there exist polynomials $k_0 : \mathbb{N} \to \mathbb{R}^+$, $q : \mathbb{N} \times \mathbb{R}^+ \to \mathbb{R}^+$ that satisfy for all $I \in \mathcal{J}$, $\omega \in \Omega$, $k \geq k_0(|I|)$,

$$\mathcal{S}(I, k, \omega) \in \mathcal{X}(I) \tag{1}$$

and for all $\varepsilon > 0$, $k > q(|I|, \varepsilon^{-1})$ (since A is closed)

$$P(\mathcal{S}(I, k, \cdot) \in A(I)) - P(X \in A(I)) < \varepsilon. \tag{2}$$

We assume w.l.o.g. that $q(m, \varepsilon) \geq k_0(m)$ for all $m \in \mathbb{N}, \varepsilon \in \mathbb{R}^+$.

Algorithm 1 Probabilistic Algorithm for Prob. k_0 is the polynomial from Def. 2 and q is the polynomial from Def. 3.

$k = \max\left\{q\left(|I|, \frac{2}{1-t}\right), k_0(I)\right\}$

choose random $\omega \in \Omega$

```
compute a sample X_k := S(I, k, ω)
if f(I, X_k) ≤ 0 then
  return NO
else
  return YES
end if
```

Lemma 1 *For a given input $I \in \mathcal{J}$ and $t \geq t(I)$ Algorithm 1 returns NO with probability at most $\frac{t+1}{2}$ if* Prob$(I) = 1$ *and it always returns NO if* Prob$(I) = 0$.

Proof **Case: There exists a $x \in \mathcal{X}(I)$ with $f(x) > 0$:**

By (Eq 2) it follows for all $k > q(|I|, \varepsilon^{-1})$ that:

$$P(f(\mathcal{S}(I,k,\cdot)) \leq 0) < t(I) + \varepsilon \leq t + \varepsilon$$

By choosing $\varepsilon = \frac{1-t}{2}$, we obtain

$$P(\mathcal{S}(I,k,\cdot) \in A) \;<\; \frac{t+1}{2}.$$

Thus, Alg. 1 will return NO although the correct answer is YES with probability at most $\frac{t+1}{2}$.

Case $f(x) \leq 0$ for all $x \in \mathcal{X}(I)$:

It follows that $f(\mathcal{S}(I, k, \omega)) \leq 0$ for all $\omega \in \Omega$, $k \geq k_0(I)$ by Def. 2. Hence, the answer of the algorithm will always be NO, if the correct answer is NO.

To prove that the problem would be in **RP**, we still have to increase the probability of YES in the positive case. This can be done by re-running the algorithm.

By Def. 4 there exists a polynomial p with $\frac{1}{1-t(I)} \leq p(|I|)$. We choose $t := \frac{p(|I|)-1}{p(|I|)}$ and it follows that $\frac{1}{1-t} = p(|\,I\,|)$ and $t(I) \leq t$. Hence, we can apply Lemma 1 without having to know $t(I)$.

By construction of Alg. 1 the computation of X_k takes time $O(q(|\,I\,|, \frac{2}{1-t}))$. We observe that the encoding for the computed sample X_k is bounded by the computation time $O(q(|\,I\,|, \frac{2}{1-t}))$. Hence, by Def. 1 there exists a polynomial g such that the runtime of Alg. 1 is bounded by $O\left(g\left(|\,I\,|, q(|\,I\,|, \frac{2}{1-t})\right)\right)$.

To obtain a correct result if the correct answer is YES with probability at least $\frac{1}{2}$, we re-run the algorithm at least $\frac{1}{\log_2\left(\frac{2}{t+1}\right)}$ times with independent choice of $\omega \in \Omega$ for each run and return YES if one of the runs returned yes.

Since the probability of NO in one run is at most $\frac{t+1}{2}$, it follows that the probability for NO in all runs is at most

$$\left(\frac{t+1}{2}\right)^{\frac{1}{\log_2\left(\frac{2}{t+1}\right)}} = 2^{\frac{\log_2\left(\frac{t+1}{2}\right)}{\log_2\left(\frac{2}{t+1}\right)}} = 2^{-\frac{\log_2\left(\frac{2}{t+1}\right)}{\log_2\left(\frac{2}{t+1}\right)}} = \frac{1}{2}.$$

We can estimate the number of iterations by observing that

$$t = \frac{p(|I|)-1}{p(|I|)} \Rightarrow \frac{2}{t+1} = \frac{2}{\frac{p(|I|)-1}{p(|I|)}+1} = \frac{2p(|I|)}{2p(|I|)-1}$$
$$\Rightarrow \frac{1}{\log_2\left(\frac{2}{t+1}\right)} = \frac{1}{\log_2\left(\frac{2p(|I|)}{2p(|I|)-1}\right)}.$$

Using the Theorem of l'Hopital we have

$$\begin{aligned}\lim_{p\to\infty} \frac{p^{-1}}{\ln\left(\frac{p}{p-1}\right)} &= \lim_{p\to\infty} \frac{p^{-1}}{\ln p - \ln(p-1)} \\ &= \lim_{p\to\infty} \frac{-p^{-2}}{\frac{1}{p}-\frac{1}{p-1}} \\ &= \lim_{p\to\infty} \frac{-p^{-2}}{\frac{p-1-p}{p(p-1)}} \\ &= \lim_{p\to\infty} p^{-2}(p^2-p) \\ &= 1\end{aligned}$$

Hence, we can bound the number of iterations by

$$\frac{1}{\log_2\left(\frac{2}{t+1}\right)} = \frac{1}{\log_2\left(\frac{2p(|I|)}{2p(|I|)-1}\right)} = O(p(|I|)).$$

Thus, we get a YES if the correct answer is YES with probability at least $\frac{1}{2}$ after a running time of

$$O\left(g\left(|I|, q\left(|I|, \frac{2}{1-t}\right)\right)\right)\frac{1}{\log_2\left(\frac{2}{t+1}\right)}$$
$$\leq O(g(|I|, q(|I|, 2p(|I|)))p(|I|)).$$

We have shown under the assumption of the existence of a sampling algorithm with the given properties that Prob is in **RP**. Since Prob is also **NP**-hard, the existence of such a sampling algorithm implies **RP** = **NP**. Hence, no such sampling algorithm can exist if **RP** $\neq$ **NP**.

Discussion

We observe that the conditions that we require for Thm. 1 on the sampling algorithm are very weak. We do not require uniform distribution, we only require that with some polynomially small probability we also sample fluxes unequal to zero in our target distribution and that we converge in polynomial time to this target distribution.

Assuming **RP** $\neq$ **NP**, it follows that for every sampling algorithm on the thermodynamic flux space there exist networks where the algorithm has one of the following properties:

- The sampling algorithm does not converge in polynomial time to the target distribution, or
- the target distribution is trivial (i.e., the probability of sampling 0 is 1).

Of course, we may be lucky and the algorithm actually samples a non-trivial distribution for the input networks. However the result says that there are networks for which the sampling algorithm will only sample 0 fluxes for some reactions and indeed, we saw that this happens also in practice not only for sampling the thermodynamically constrained flux space but also the ordinary steady-state flux space. It might be a property of the ACHR method that even for the ordinary steady-state flux space only 0 fluxes for some reactions are sampled, because in theory classical hit and run sampling (with appropriate rounding to remove the heterogeneous scales of metabolic networks) is guaranteed to sample uniformly [1]. De Martino et al. [43] showed that indeed ACHR seems to have problems with high-dimensional instances, like 500 dimensional uniform hypercubes. Other sampling methods, e.g. loopy-belief propagation [10, 11] or poling-based methods [9] might not have these problems.

Unreliable sampling is very critical, since we then may be led to the false assumption that the reaction is never used, although it actually could be. To make sure that such results are true, it is essential to verify them with a deterministic method. In the case of deciding whether flux is possible through a given reaction, we can decide this by solving an optimization problem [21, 37].

We have shown that sampling artifacts happen for the flux variability problem with thermodynamic constraints (and in practice they even happen without thermodynamic constraints with ACHR sampling). However, sampling is used to check a wide variety of different properties. Although the result does not directly imply that sampling results for these other properties are unreliable as well, caution is highly advised. For example, consider correlation / flux coupling analysis [20]. If a reaction always carries zero flux in all samples by an artifact, although it can also carry non-zero flux, it follows that this reaction seems uncorrelated to all other reactions. However, it may very well be correlated / coupled. In Fig 1, we see such an example. Assume the flux space (see Fig 2) is sampled using a uniform distribution. Then, we will almost surely never sample non-zero flux through reaction r_1. Correlation analysis would yield that flux through r_1 is uncorrelated (they are even independent) to flux through r_2 and r_3, although the fluxes are actually exclusive (e.g. r_1 and r_2 cannot carry flux at the same time). On the other hand, if reaction r_2 would be removed (because it is simply the aggregation of r_3 and r_4) the result would change significantly. Then, with uniform sampling, positive fluxes through r_1 and through r_3 would be sampled with equal probability.

Conclusion

Although sampling has been used successfully for the analysis of many different kind of problems, the results obtained by sampling should be used with caution. In the field of metabolic network analysis in particular, we showed (assuming **RP** $\neq$ **NP**) that for every polynomial-time sampling method obeying thermodynamic constraints there exist networks for which the sampling method will produce artifacts. Hence, qualitative results obtained by sampling complex spaces should always be double checked by a different method.

Supporting Information

S1 Code. Source code. Matlab scripts used to compute the results in the manuscript.
(ZIP)

Author Contributions

Conceived and designed the experiments: ACR. Performed the experiments: ACR. Analyzed the data: ACR. Wrote the paper: ACR.

References

1. Lovász L. Hit-and-run mixes fast. Mathematical Programming. 1999; 86(3):443–461.
2. Dyer M, Frieze A, Kannan R. A random polynomial-time algorithm for approximating the volume of convex bodies. Journal of the ACM. 1991; 38(1):1–17. doi: 10.1145/102782.102783
3. Lovász L, Vempala S. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. Journal of Computer and System Sciences. 2006; 72(2):392–417. doi: 10.1016/j.jcss.2005.08.004
4. Elekes G. A geometric inequality and the complexity of computing volume. Discrete and Computational Geometry. 1986; 1(1):289–292. doi: 10.1007/BF02187701
5. Bárány I, Füredi Z. Computing the volume is difficult. Discrete and Computational Geometry. 1987; 2(1):319–326. doi: 10.1007/BF02187886
6. Schuster S, Hilgetag C. On elementary flux modes in biochemical systems at steady state. J Biol Systems. 1994; 2:165–182. doi: 10.1142/S0218339094000131
7. Schilling CH, Letscher D, Palsson BO. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function form a pathway-oriented perspective. Journal of Theoretical Biology. 2000; 203:229–248. doi: 10.1006/jtbi.2000.1073 PMID: 10716907
8. Schellenberger J, Palsson BO. Use of randomized sampling for analysis of metabolic networks. The Journal of Biological Chemistry. 2009; 284(9):5457–5461. doi: 10.1074/jbc.R800048200 PMID: 18940807
9. Binns M, de Atauri P, Vlysidis A, Cascante M, Theodoropoulos C. Sampling with poling-based flux balance analysis: optimal versus sub-optimal flux space analysis of Actinobacillus succinogenes. BMC Bioinformatics. 2015; 16:49. doi: 10.1186/s12859-015-0476-5 PMID: 25887116
10. Braunstein A, Mulet R, Pagnani A. Estimating the size of the solution space of metabolic networks. BMC Bioinformatics. 2008; 9:240. doi: 10.1186/1471-2105-9-240 PMID: 18489757
11. Font-Clos F, Massucci FA, Castillo IP. A weighted belief-propagation algorithm for estimating volume-related properties of random polytopes. Journal of Statistical Mechanics: Theory and Experiment. 2012;. doi: 10.1088/1742-5468/2012/11/P11003
12. Kaleta C, de Figueiredo LF, Behre J, Schuster S. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In: Grosse I, Neumann S, Posch S, Schreiber F, Stadler P, editors. Lecture Notes in Informatics—Proceedings. vol. P-157. Gesellschaft für Informatik; 2009. p. 179–189.
13. Thiele I, Price ND, Vo TD, Palsson BO. Candidate Metabolic Network States in Human Mitochondria. The Journal of Biological Chemistry. 2005; 280:11683–11695. doi: 10.1074/jbc.M409072200 PMID: 15572364
14. Bordbar A, Lewis NE, Schellenberger J, Palsson BO, Jamshidi N. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. Molecular Systems Biology. 2010; 6:422. doi: 10.1038/msb.2010.68 PMID: 20959820
15. Machado D, Soons Z, Patil KR, Ferreira EC, Rocha I. Random sampling of elementary flux modes in large-scale metabolic networks. Bioinformatics. 2012; 28:i515–i521. doi: 10.1093/bioinformatics/bts401 PMID: 22962475
16. Almaas E, Kovács B, Vicsek T, Oltvai ZN, Barabási AL. Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature. 2004; 427:839–843. doi: 10.1038/nature02289 PMID: 14985762
17. Papin JL, Reed Jason A, Palsson BO. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. TRENDS in Biochemical Sciences. 2004; 29(12):641–647. doi: 10.1016/j.tibs.2004.10.001 PMID: 15544950
18. Sarıyar B, Perk S, Akman U, Hortaçsu A. Monte Carlo sampling and principal component analysis of flux distributions yield topological and modular information on metabolic networks. Journal of Theoretical Biology. 2006; 242:389–400. doi: 10.1016/j.jtbi.2006.03.007 PMID: 16860341
19. Kelk SM, Olivier BG, Stougie L, Bruggeman FJ. Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. Scientific Reports. 2012; 2:580. doi: 10.1038/srep00580 PMID: 22896812
20. Xi Y, Chen YPP, Chen Q, Wang F. Comparative study of computational methods to detect the correlated reaction sets in biochemical networks. Briefings in Bioinformatics. 2011; 12(2):132–150. doi: 10.1093/bib/bbp068 PMID: 20056730
21. Schellenberger J, Lewis NE, Palsson BØ. Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophysical Journal. 2011; 100:544–553. doi: 10.1016/j.bpj.2010.12.3707 PMID: 21281568
22. Beard DA, Babson E, Curtis E, Qian H. Thermodynamic constraints for biochemical networks. Journal of Theoretical Biology. 2004; 228:327–333. doi: 10.1016/j.jtbi.2004.01.008 PMID: 15135031

23. Kaufman DE, Smith RL. Direction choice for accelerated convergence in hit-and-run sampling. Operations Research. 1998; 46(1):84–95. doi: 10.1287/opre.46.1.84

24. Papadimitrou CH. Computational Complexity. Addison-Wesley; 1994.

25. Beard DA, dan Liang S, Qian H. Energy balance for analysis of complex metabolic networks. Biophysical Journal. 2002; 83:79–86. doi: 10.1016/S0006-3495(02)75150-3 PMID: 12080101

26. Beard DA, Qian H. Thermodynamic-based computational profiling of cellular regulatory control in hepatocyte metabolism. American Journal of Physiology—Endocrinology and Metabolism. 2005; 288: E633–E644. doi: 10.1152/ajpendo.00239.2004 PMID: 15507536

27. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular Systems Biology. 2007; 3:121. doi: 10.1038/msb4100155 PMID: 17593909

28. Fleming RMT, Thiele I, Nasheuer HP. Quantitative assignment of reaction directionality in constraint-based models of metabolism: Application to *Escherichia coli*. Biophysical Chemistry. 2009; 145:47–56. doi: 10.1016/j.bpc.2009.08.007 PMID: 19783351

29. Jol SJ, Kümmel A, Terzer M, Stelling J, Heinemann M. System-Level Insights into Yeast Metabolism by Thermodynamic Analysis of Elementary Flux Modes. PLoS Computational Biology. 2012; 8:3. doi: 10.1371/journal.pcbi.1002415

30. Kümmel A, Panke S, Heinemann M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Molecular Systems Biology. 2006; 2:2006.0034. doi: 10.1038/msb4100074 PMID: 16788595

31. Kümmel A, Panke S, Heinemann M. Systematic assignment of thermodynamic constraints in metabolic network models. BMC Bioinformatics. 2006; 7:512. doi: 10.1186/1471-2105-7-512 PMID: 17123434

32. Hoppe A, Hoffmann S, Holzhütter HG. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. BMC Systems Biology. 2007; 1:23. doi: 10.1186/1752-0509-1-23 PMID: 17543097

33. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. Biophysical Journal. 2006; 90(4):1453–1461. doi: 10.1529/biophysj.105.071720 PMID: 16299075

34. Singh A, Soh KC, Hatzimanikatis V, Gill RT. Manipulating redox and ATP balancing for improved production of succinate in E. coli. Metabolic Engineering. 2011; 13:76–81. doi: 10.1016/j.ymben.2010.10.006 PMID: 21040799

35. Price ND, Thiele I, Palsson BO. Candidate states of *Helicobacter pylori's* genome-scale metabolic network upon application of "Loop Law" thermodynamic constraints. Biophysical Journal. 2006; 90: 3919–3928. doi: 10.1529/biophysj.105.072645 PMID: 16533855

36. Cogne G, Rügen M, Bockmayr A, Titical M, Dussap CG, Cornet JF, et al. A model-based method for investigating bioenergetic processes in autotrophically growing eukaryotic microalgae: Application to the green algae *Chlamydomonas reinhardtii*. Biotechnol Progress. 2011; 27(3):631–640. doi: 10.1002/btpr.596

37. Müller AC, Bockmayr A. Fast Thermodynamically constrainted Flux Variability Analysis. Bioinformatics. 2013; 29(7):903–909. (AC Müller is now called AC Reimers). doi: 10.1093/bioinformatics/btt059 PMID: 23390138

38. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering. 2003; 5:264–276. doi: 10.1016/j.ymben.2003.09.002 PMID: 14642354

39. Gudmundsson S, Thiele I. Computationally efficient flux variability analysis. BMC Bioinformatics. 2010; 11:489. doi: 10.1186/1471-2105-11-489 PMID: 20920235

40. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols. 2011; 6(9):1290–1307. doi: 10.1038/nprot.2011.308 PMID: 21886097

41. Schellenberger J, Park JO, Conrad TM, Palsson BO. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics. 2010; 11:213. doi: 10.1186/1471-2105-11-213 PMID: 20426874

42. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism. Molecular Systems Biology. 2011; 7:535. doi: 10.1038/msb.2011.65 PMID: 21988831

43. De Martino D, Mori M, Parisi V. Uniform Sampling of Steady States in Metabolic Networks: Heterogeneous Scales and Rounding. PLoS ONE. 2015; 10(4):e0122670. doi: 10.1371/journal.pone.0122670 PMID: 25849140