Best practices for reporting throughput in biomedical research

To the Editor — In the last few decades, technological advances in photonics, electronics, computing, fluidics, robotics and chemistry have substantially boosted the rates of data acquisition and processing and the scale of automation and parallelization. This has enabled high-throughput performance in measurement, imaging, screening, sequencing, manipulation and sorting of molecules, compounds, genes and cells[1–6]. The term “throughput” is widely accepted and constantly used in the biomedical community, where high-throughput operations are indispensable for efficient, reproducible, time-sensitive, low-cost and rare-event applications.

Unfortunately, the term is often vaguely defined and inconsistently used in different domains of the life sciences. In fact, it is used differently from the rigorously defined “throughput” in the field of electrical engineering, where throughput expresses the maximum amount of data that can theoretically be sent or processed in a given amount of time and is quantified in units of bits per second[7]. In biomedical settings (in particular, high-throughput flow cytometry[1], screening[2] and sequencing[5,6]), where the throughput depends on non-electrical parameters (for example, chemical, biological or fluidic) that are often non-uniform and time-varying, the practically achievable throughput can be orders of magnitude lower than the theoretical throughput. Furthermore, different biomedical domains employ different units of throughput, contributing to confusion and hindering collaboration between the domains[1,8]. These problems are troublesome, but their importance is often neglected.

To better understand the problems and rectify them, we revisit a general equation (Fig. 1a) that defines the throughput T of a high-throughput operation using one or more high-throughput instruments (for example, a flow cytometer, a DNA sequencing machine, a high-throughput screening system or a combination of these)[7]:

$$T = \frac{\sum_{i=1}^{N} n_i D_i}{\sum_{i=1}^{N} \tau_i} \qquad (1)$$

where N is the total number of samples or steps per run, n_i is the number of discrete objects in sample i that are processed (for example, molecules, compounds, genes, cells or organoids that are measured, imaged, screened, sequenced, manipulated or sorted), D_i is the depth of data about each object (for example, the number of bits encoded by a fluorescence signal, an image or a mass spectrum) used when handling sample i, and τ_i is the duration of time required for handling sample i and is given by the sum of sample preparation and loading time τ_i^a, sample processing time τ_i^b, and sample removal and recovery time τ_i^c; that is, $\tau_i = \tau_i^a + \tau_i^b + \tau_i^c$. If n_i, τ_i

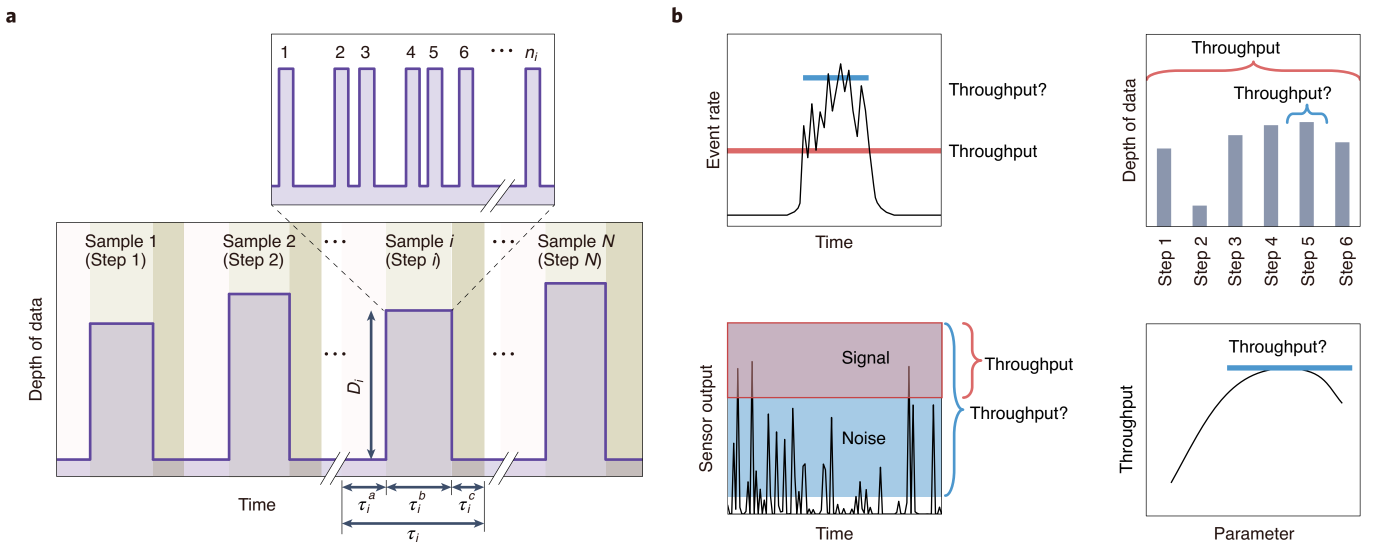

Fig. 1 | Throughput definition and common pitfalls. a, Definition of throughput. **b**, Common pitfalls and solutions. Top left: instead of omitting the sample preparation or loading and sample removal or recovery time durations and reporting the momentary maximum of event rate (blue), the global average throughput including the sample preparation or loading and sample removal or recovery time durations (red) should be reported. Meanwhile, the computational times of postprocessing analyses can be omitted since they are not directly related to the raw-data acquisition. If the sample preparation or loading and sample removal or recovery time durations are omitted from the throughput calculation, the time durations should be reported as well. Top right: instead of reporting the throughput of a chosen step (blue), all steps of the entire high-throughput operation and their depths of data need to be considered to determine the throughput red). Bottom left: before determining throughput, it is necessary to define which event corresponds to a target event. Instead of using all events, including noise events (blue), only the count of target events (red) should be used to determine throughput. Bottom right: black curve indicates the dependency of throughput on another parameter. Conditions for achieving the maximum throughput (blue) should be reported together with the throughput (red).

and D_i are identical for all samples and objects, then equation (1) can be simplified to $T = Nn_1D_1/\tau_1$. Since the data about each object are electrically acquired and digitally processed in modern scientific instrumentation, the throughput is expressed in bits per second as in electrical engineering, regardless of the type of data. Note that the throughput can be increased even at a low event rate if the depth of data is large. See Supplementary Note 1 for exemplary computations for imaging flow cytometry and high-throughput screening.

Common pitfalls are illustrated in Fig. 1b. First, the most frequent issue is that parts in equation (1) are omitted when calculating throughput. For example, in flow cytometry, it is commonly assumed that $N = 1$, $\tau_1^a \ll \tau_1^b$ and $\tau_1^c \ll \tau_1^b$, such that equation (1) reduces to $T \approx n_1D_1/\tau_1^b$, where n_1/τ_1^b (known as the event rate) is directly proportional to and often treated as throughput although it is not entirely accurate. Furthermore, the highest event rate may only be achieved temporarily, such that the average event rate $\langle \mathrm{d}n_1/\mathrm{d}t \rangle$ over τ_1^b is lower than the momentary event rate $\mathrm{d}n_1/\mathrm{d}t$, but they are often regarded as equal[9]. Second, since a high-throughput operation is often composed of multiple steps, especially in high-throughput screening, a typical pitfall is that only the throughput of a chosen step is reported, with the rate-limiting step ignored. Also, in high-throughput screening, the depth of data is often overlooked, with only the number of wells per day considered even though readouts have progressed from simple plate-reader-based assays to complex microscopy images[10]. Third, the throughput can be inflated by including 'non-usable' events (for example, measurement noise, cell debris, cell doublets, Poisson-limited multi-cell compartmentalization in droplets, and non-targeted reads) into the event count n_i. Fourth, it is misleading to treat throughput as a constant parameter and ignore its dependency on other parameters (for example, sample concentration, sample delivery speed, sequence diversity, fragment size, or the response time and success rate of a high-throughput operation).

These pitfalls can be avoided by recognizing equation (1). In electrical engineering, throughput is assessed in a more differentiated manner by providing the theoretically possible maximum throughput and the practically achievable throughput (called "goodput"[7]). These two definitions can provide guidelines to solve the above pitfalls. First, as throughput is a measure for a continuous process in which the sample preparation or loading and removal or recovery are involved, all these parts should be considered to calculate the throughput. If the event rate is reported instead of T, τ_1^a and τ_1^c should also be reported. Second, T needs to take all steps of the high-throughput operation and their depths of data into consideration as defined in equation (1). Third, to exclude the contribution of non-usable events, it is essential to provide a definition of target objects. If the portion of non-usable events is larger (at least a few times) than that of good events, the goodput should be calculated by excluding non-usable events in postprocessing steps from the total event count. Fourth, since throughput is often a function of several other parameters, their values should ideally be measured and reported under different conditions, or at least the conditions on the throughput should be reported if the reported throughput is a primary point of novelty claimed.

Maik Herbig[1,17], Akihiro Isozaki[1,17], Dino Di Carlo[2,3,4], Jochen Guck[5,6], Nao Nitta[7], Robert Damoiseaux[2,3,8], Shogo Kamikawaji[9], Eigo Suyama[9], Hirofumi Shintaku[10], Angela Ruohao Wu[11,12], Itoshi Nikaido[13,14,15] and Keisuke Goda[1,2,16] ✉

[1]Department of Chemistry, The University of Tokyo, Tokyo, Japan. [2]Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA, USA. [3]California NanoSystems Institute, University of California, Los Angeles, Los Angeles, CA, USA. [4]Department of Mechanical Engineering, University of California, Los Angeles, Los Angeles, CA, USA. [5]Max Planck Institute for the Science of Light, Erlangen, Germany. [6]Max-Planck-Zentrum für Physik und Medizin, Erlangen, Germany. [7]CYBO, Inc., Tokyo, Japan. [8]Department of Molecular and Medicinal Pharmacology, University of California, Los Angeles, Los Angeles, CA, USA. [9]Chugai Pharmaceutical Co., Ltd., Tokyo, Japan. [10]RIKEN Cluster for Pioneering Research, Saitama, Japan. [11]Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China. [12]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. [13]RIKEN Center for Biosystems Dynamics Research, Saitama, Japan. [14]Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. [15]Graduate School of Science and Technology, University of Tsukuba, Saitama, Japan. [16]Institute of Technological Sciences, Wuhan University, Wuhan, China. [17]These authors contributed equally: Maik Herbig, Akihiro Isozaki.

✉*e-mail: goda@chem.s.u-tokyo.ac.jp*

Published online: 04 May 2022
https://doi.org/10.1038/s41592-022-01483-6

References

1. Robinson, J. P., Rajwa, B., Patsekin, V. & Davisson, V. J. *Expert Opin. Drug Discov.* **7**, 679–693 (2012).
2. Macarron, R. et al. *Nat. Rev. Drug Discov.* **10**, 188–195 (2011).
3. Nawaz, A. A. et al. *Nat. Methods* **17**, 595–599 (2020).
4. Nitta, N. et al. *Cell* **175**, 266–276.e13 (2018).
5. Reuter, J. A., Spacek, D. V. & Snyder, M. P. *Mol. Cell* **58**, 586–597 (2015).
6. Loman, N. J. et al. *Nat. Biotechnol.* **30**, 434–439 (2012).
7. Comer, D. *Computer Networks and Internets* (Pearson, 2016).
8. Parola, C., Neumeier, D. & Reddy, S. T. *Immunology* **153**, 31–41 (2018).
9. Di Carlo, D. et al. *Science* **364**, eaav1429 (2019).
10. Boutros, M., Heigwer, F. & Laufer, C. *Cell* **163**, 1314–1325 (2015).

Acknowledgements

This work is supported by JSPS Core-to-Core Program, JSPS KAKENHI grant numbers 19H05633 and 20H00317, JSPS-NUS/NTU Bilateral Joint Research Program 2021, AMED grant numbers JP20wm0325021 and 21ak0101163h0001, the White Rock Foundation, the Ogasawara Foundation, the Nakatani Foundation, the Konica Minolta Foundation and the Chemo-Sero-Therapeutic Research Institute.

Author contributions

K.G. conceived the project. M.H. prepared the initial manuscript with the support of K.G. and A.I. All authors revised and edited the manuscript. K.G. acquired funding.

Competing interests

K.G. and N.N. are shareholders of CYBO. All remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01483-6.

Peer review information *Nature Methods* thanks Kevin Tsia and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.