IEEE

CDS NEWSLETTERS

Volume 18, Number 4 ISSN 1550-1914 November 2024

Development of Natural and Artificial Intelligence

Invalidity of the Experimental Protocol in Two Nobel Prizes

Juyang Weng, Brain-Mind Institute and GENISAMA, USA, Email: jweng@genisama.com
Colin Schmidt, Le Mans University, France, Email: colin.schmidt@univ-lemans.fr
Dongshu Wang, Zhengzhou University, China, Email: wangdongshu@zzu.edu.cn
Ming Xie, Nanyang Technological University, Singapore, Email: mmxie@ntu.edu.sg

ABSTRACT: *For discussions in the scientific community, we raise concerns over the Post-Selection protocols in the Nobel Prize for Physics 2024 and the Nobel Prize for Chemistry 2024, as well as the current flood of Post-Selection in Artificial Intelligence (AI) and machine learning. We hypothesize that the Post-Selection protocol has three mistakes: (a) missing a test, (b) hiding bad-looking data, and (c) exaggerating prediction accuracy. However, the future of AI and machine learning seems to be bright. This is a preprint.*

KEYWORDS: Nobel Prizes, Research Ethics, Experimental Protocol, Post-Selection, Performance Evaluation, Machine Learning, Artificial Intelligence, Local Minima, Developmental Networks, Maximum Likelihood

Every year, the Nobel Committees perform extremely challenging work to select probably the most known awards mankind has known. Naturally, the members of the Nobel Committees have been conscientious and conservative. For example, Albert Einstein, the "person of the century" named by the *Time* magazine, is best known for his contributions to the theory of relativity. However, he did not receive a Nobel for relativity. Instead, the Nobel Committee for Physics 1921 gave him a Nobel for "photoelectric effect". We applaud the Nobel Committees for their hard work. However, when an invalid protocol has flooded machine learning and the media, even the careful and conservative Nobel Committees are vulnerable to blunders, especially after the establishment (e.g., *Science*, *Nature*, other journals and large corporations) has purposefully ignored reports about the invalidity of Post-Selection protocol [1, 2]. The pressure to publish or perish should have also played a role in neglecting the literature.

In early October this year, the Nobel Prize for Physics and the Nobel Prize for Chemistry went to machine learning works. While we are pleased to see the physics and chemistry disciplines recognize the influence of machine learning, here we raise serious concerns over the experimental protocol used by these two awards and the flood of such protocol in AI and other sciences. For example, almost all machine learning manuscripts received by the *International Journal of Humanoid Robotics* employed such a protocol.

Although this protocol has different names in statistics, we call it Post-Selection here. Normally, an experimenter selects a model *before* the model applies to the random data. As an analogy, an experimenter writes a lottery ticket (model) *before* the lottery organizer goes through a random draw across a large number *n* of lottery tickets (models). In contrast, in the Post-Selection protocol, the experimenter selects the luckiest model *after* (i.e., the term *Post* in Post-Selection) all the systems have applied to a random data set called the validation set V. By validation set, we mean the chooser knows the set. In our lottery analogy, the experimenter reports only *after* the luckiest lottery ticket is known in the published lottery draw and hides all other less-lucky tickets.

It is well-known that the initial states of a system greatly affect the final accuracy of the system prediction [3, 4, 5, 6]. Even the tightly structured Hidden Markov Model needs to be initialized by pre-segmentation, e.g., using a separate K-means clustering process [7, p.274], although the sensitivity of K-means to initial segmentation is also a well-recognized problem. Some highly constrained linear models, such as the Principal Component Analysis [8] may still lead to different representations depending on their initial eigenvectors, especially when multiple eigenvalues are similar. The Post-Selection protocol dates back at least to the Proprotional-Integral-Derivative (PID) controller, but the PID controller is not comparable with the two Nobel Prizes in terms of the number of model parameters (e.g., 3 in PID but 60 million in [4]) and the number of local minima.

As correctly stated by Berk et al. [9], “a data-driven variable selection process produces a model that is itself stochastic”. The statistic (e.g., prediction error) of the luckiest model is stochastic, which is a function of all n trained models (due to the Post-Selection). Cross-validation (such as the leaving-one-batch-out procedure) by using different splits of the input data set D (into disjoint F and V so that $D = F \cup V$) only reduces the bias of a particular split; but cross-validation does not eliminate the nature that each luckiest model depends on all n the candidate models [10].

When the number of free parameters is small compared with the number of parameters that generate the random data, e.g., synthetic data cases in [10] or linear models in [9], the luckiest model is only an overfit of a particular validation set V.

However, in AI, the dimension of free parameters is extremely high, e.g., 60 million parameters in [4]. The luckiest model on any validation set V can be perfect (zero error) as proven in the theory of the Pure-Guess Nearest Neighbor (PGNN) model [2]. (The Nearest Neighbor With Threshold (NNWT) model cited in [2] is more efficient than PGNN by guessing output only when the current input is far from all data in the fit set F.) In other words, if one buys enough lottery tickets, he can almost be sure to get a ticket that hits the jackpot.

Therefore, the post-selected luckiest accuracy in AI is meaningless, because no system can beat PGNN and NNWT in prediction accuracy. Why? The luckiest PGNN and NNWT model gives a perfectly zero error for any given validation set V. Both PGNN and NNWT store the entire set F. NNWT generalizes from F to V using Euclidean distance generalization if the input is not too far from the nearest neighbor in F. Otherwise, NNWT just guesses an output. PGNN always uses its Pure Guess (PG) regardless of how far the input is.

Fig. 1 illustrates the Post-Selection process. The luckiest random ball among $n = 3$ balls has the luckiest Post-Error on the validation set V, but it does not tell anything about the test error (green curve) because a test set T is absent. This absence should be true for two Nobel Prizes (e.g., see AlphaFold2 below for more detail).

The technical aspects of why Post-Selection is misconduct have been covered in [1, 2, 11]. Here is a summary. Suppose that a human trains n models each of which starts from a different set of hyper-

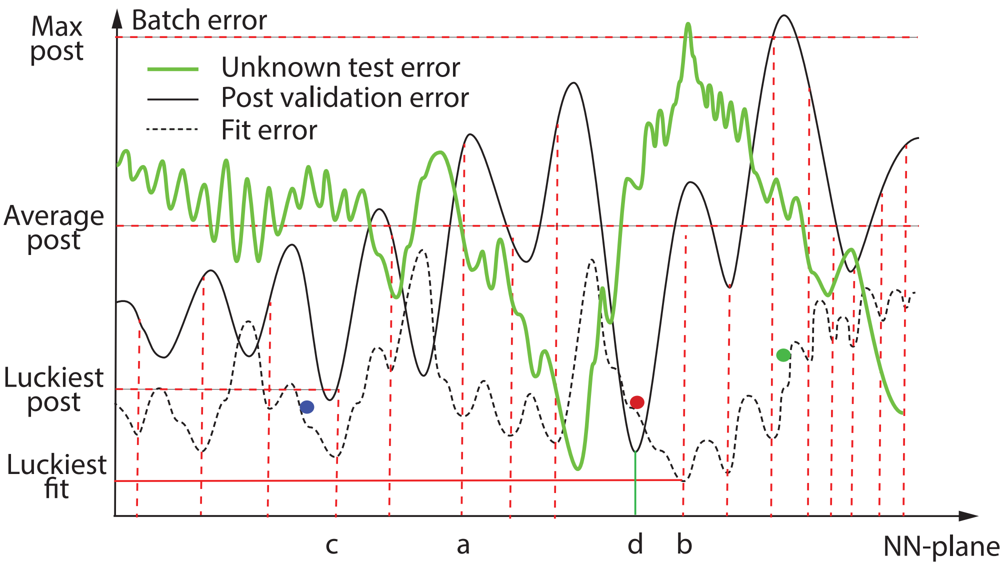

Figure 1: A 1D-terrain illustration for the invalid Post-Selection protocol. The dashed terrain is from the fit set F; the solid-black terrain is from the validation set V, and the green terrain is from a test set T that does not exist in post-selection. Many iterative learning processes drop random balls (3 here) that land on a random location on the dashed terrain and then roll down the hill until they get stuck into a local pit. The luckiest model is the blue ball whose validation height (distance to the solid-black curve) is the lowest (marked as Luckiest post) among the three balls. However, the luckiest ball has not been tested (unknown green terrains) and all other balls should be transparently reported about their random distribution on validation of V. The NN-plane is the parameter space of a model, e.g., 60-million dimension in [4]. Courtesy of [11].

parameters and neuronal weights. These n models fit a fit set F, using a greedy fitting method, such as so-called "deep learning" based on supervised learning, reinforcement learning, adversarial learning, etc. The human validates the n models using a validation set V, and computed the validation error e_i, $i = 1, 2, ..., n$. He must report a single error $\hat{e}$ that should have the minimum mean square error (MMSE) from the n errors e_i, $i = 1, 2, ..., n$:

$$\hat{e} = \underset{\hat{e}}{\operatorname{argmin}} \sum_{i=1}^{n} (\hat{e} - e_i)^2 P_i \tag{1}$$

where P_i is the probability of observing e_i. Assuming that all e_i, $i = 1, 2, ..., n$, have the same probability, Theorem 1 and its proof in [11] have established the MMSE estimate of $\hat{e}$ is the average of the n errors of all n trained models:

$$\hat{e} = \frac{1}{n} \sum_{i=1}^{n} e_i. \tag{2}$$

The Post-Selection misconduct reports the luckiest (smallest) e_i on the validation set V (or test set T). Without loss of generality, suppose the n errors are ranked in nondecreasing order:

$$e_1 \leq e_2 \leq ... \leq e_n. \tag{3}$$

Instead of reporting the MMSE estimate $\hat{e}$ in Eq. (2), Post-Selection reports the luckiest e_1. For example, for the MNIST data set prediction in [12], the luckiest $e_1 = 1.58\%$, while the average $\hat{e} = 15.0\%$ which is $\hat{e}/e_1 = 9.5$ times larger than the luckiest e_1.

The Post-Selection protocol contains three types of misconduct:

(a) Missing a test: Contrary to using the "test" word [3, 4, 5, 6], there is an absence of a test—like the luckiest lottery ticket in the past without a future test. This means all so-called "applications" in deep learning research only fit a validation set, without "applications".

(b) Hiding bad-looking data: There is a lack of transparency about the error distribution of all n trained models [3, 4, 5, 6]—like hiding all loser lottery tickets. This means "using the validation set for post-selection during model selection and hyperparameter tuning" is a common practice that has only negative values.

(c) Exaggerating prediction accuracy: The validation accuracy reported [3, 4, 5, 6] is based on only the luckiest model and thus, is inflated (e.g., 9.5 times in the above example [12]). Instead, the validation accuracy should be MMSE, i.e., the average of all n trained models—like the average return rate of all lottery tickets, every lottery ticket having the same expected return rate in the future.

For example, the ImageNet Contests [13] and the CASP protein folding contests [14] appear to have greatly exaggerated the prediction accuracies through organizer-facilitated Post-Selections. The organizers reported only the luckiest one among $n \geq 5$ submissions from each group.

ImagetNet allowed any finite number n of submissions from each group by providing a so-called online "Evaluation Server". Namely, the Post-Selection protocol was hidden in the ImageNet competition [13] since each group could conduct Post-Selection using the "Evaluation Server". Even the test set T was released publicly—"A set of test images is also released" [13] with only partial annotations withheld. However, each group could manually supplement the remaining annotations or, alternatively, guess them and then check the "Evaluation Server". A piece of evidence of post-selection using test set is that for the luckiest network, its error (15.3%) on the test set is even smaller than its error (15.4%) on the validation set [13, Table 2]. The luckiest fitting accuracies (which are not test errors) of SuperVision [4] reported by the ImageNet [13] seem to have fooled the Turing Award Committee 2018 and the Nobel Committee for Physics 2024.

Likewise, the AlphaFold2 fitting accuracies (which are not test errors) in protein folding contest CASP14 [6] seem to have fooled the Nobel Committee for Chemistry 2024. For CASP14, $n = 5$ but the Post-Selection protocol is hidden by the CASP14 organizer. John Moult, the contact author of the contest organizer's paper [14] privately emailed to Juyang Weng, "up to five models each group is allowed to submit on one domain", apparently meaning $n = 5$ in Post-Selection. The authors are still waiting for John Moult to provide the 5 accuracy values of AlphaFold2. In [14] the organizer vaguely mentioned: "The set of five submitted models contain two different conformations for this region", apparently out of context about the experimental protocol. Additionally, this sentence means that Fig. 2 in [14] is an example that the 5 AlphaFold2 models are different, but the paper did not tell us how large the differences are. In other words, the Post-Selection was performed by the CASP14 organizer but the organizer hid it.

A reviewer of this article might require the authors to provide evidence of Post-Selection but these charged papers all hide the misconduct protocol. Indeed, very few experimetnal papers openly discussed the Post-Selection protocol, e.g., $n = 20$ in [15, Fig.5], $n = 20$ in [12, Fig. 7], $n = 10{,}000$ in [16, p. 232], and

$n = 30$ in [17, Fig. 2], and . Many experimental papers simply hid the Post-Selection protocol.

Furthermore, the luckiest model is a function of all n trained models. For example, in AlphaFold2, the luckiest model is a function of 5 models. Apparently, the CASP14 organizer allowed $n = 5$ models from AlphaFold2 and the organizer did the Post-Selection for AlphaFold2. After seeing the validation set V (not a test set), the CASP14 organizer declared that there is a lucky model that falls around the 1 $\overset{\circ}{A}$ ball of the particular validation set V (consisting of 87 sequences), but there are $5-1=4$ other trained AlphaFold2 models that were worse. However, the CASP14 organizer did not report the remaining 4 AlphaFold2 models as they should. Furthermore, this luckiest model has not been tested by a disjoint test set T yet. According to the above MMSE theorem, the luckiest AlphaFold2 is expected to produce the average of validation errors of all 5 AlphaFold2 models instead. Analogously, the luckiest lottery ticket in last week's lottery draw has not been tested in the future unknown lottery draw and it is expected to be not as lucky as last week—only giving average $\hat{e}$.

Therefore, all the reported prediction accuracies that use Post-Selection protocols, including those from the Hopfield Nets [3], SuperVision [4] in the ImageNet Contests, AlphaFold [6] in the CASP14 Contests, along with those in many other well-known systems (e.g., Transformers [5] and ChatGPT [18]) should have been greatly exaggerated and invalid.

Regardless of the current Post-Selection flood, the future of AI and machine learning appears to be bright. There is a holistic solution to 20 million-dollar problems [19]. What the Post-Selection protocol suffers from is Problem 13 of the 20 million-dollar problems, namely the high-dimensional local minima problem. The holistic solution trains only $n = 1$ model [1, Eq. (11)] but it is further globally optimal in the sense of distributed maximum likelihood, without any iterative search. For example, the solution must include automatic brain patterning [20] and lifelong developmental changes in the internal machinery brought on by learning itself [21]. However, this subject is beyond the scope of this short editorial.

In summary, the Post-Selection protocol should be invalid for any model that has so many parameters that it can simply overfit any given validation set through Post-Selection, like PGNN and NNWT. This protocol blunder could be also true with all awardees of the Nobel Prize for Physics 2024 and the Nobel Prize for Chemistry 2024. In particular, the so-called "successfully implement examples of deep and dense networks" and the so-called "most monomeric protein structures can now be predicted with high fidelity", respectively claimed by the two "Scientific Background" documents from the two Nobel Committees, seem to lack a valid protocol basis.

The two Nobel Committees should know what to do to live up to the public trust.

ACKNOWLEDGEMENTS: The authors would like to thank the following people who have read the paper and made comments or feedbacks: Yoshua Bengio, Peter Brzezinski, Andrei Chabes, Ulf Danielsson, Alex Graves, Demis Hassabis, Geoffrey Hinton, Alan E. Hubbard, John Hopfield, Steven Grossberg, Anders Irback, Goran Johansson, John Jumper, Nikola Kasabov, Haibo Li, Yann LeCun, Heiner Linke, Jay McClelland, Dagmar Monett-Diaz, John Moult, Eva Olsson, Mark Pearce, Alex Pentland, Giulio Sandini, Juergen Schmidhuber, Zhongzhi Shi, Linda Smith, Peter Somfai, Mriganka Sur, Jochen Triesch, Philip Walther, Deliang Wang, Xiang Wu, and Michael Zock.

References

[1] J. Weng. On post selections using test sets (PSUTS) in AI. In *Proc. Int'l Joint Conference on Neural Networks*, pages 1–8, Shenzhen, China, July 18-22 2021. NJ: IEEE Press.

[2] J. Weng. On "deep learning" misconduct. In *Proc. 2022 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 1–8, Beijing, China, Dec. 9-11 2022. SciTePress. arXiv:2211.16350.

[3] D. Krotov and J. J. Hopfield. Dense associative memory for pattern recognition,. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, pages 1–9, Long Beach, CA, 2016. NIPS Foundation.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[5] A. Vaswani, N. Shazeer, I. Polosukhin, et al. Attemtion is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 1–15, Long Beach, CA, Dec. 4-9, 2017. NIPS Foundation.

[6] J. Jumper, D. Hassabis, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[7] L. R. Rabiner, L. G. Wilpon, and F. K. Soong. High performance connected digit recognition using hidden Markov models. *IEEE Trans. Acoustics, Speech and Signal Processing*, 37:1214–1225, Aug. 1989.

[8] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[9] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[10] M. J. van der Laan, Eric C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007.

[11] J. Weng. Misconduct in post-selection and deep learning. In *Proc. the 8th International Conf. on Control, Robotics and Sybernetics*, pages 1–9, Changsha, China, Dec. 22-24 2023. NJ: IEEE Press. arXiv 2403.00773.

[12] Q. Gao, G. A. Ascoli, and L. Zhao. BEAN: Interpretable and efficient learning with biologically-enhanced artificial neuronal assembly regularization. *Front. Neurorobot*, 15:1–13, June 1 2021. https://doi.org/10.3389/fnbot.2021.567482.

[13] O. Russakovsky, J. Deng, L. Fei-Fei, et al. ImageNet large scale visual recognition challenge. *Int'l Journal of Computer Vision*, 115:211–252, 2015.

[14] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins*, 89(12):1607–1617, 2021. doi:10.1002/prot.26237.

[15] A. Graves, G. Wayne, M. Reynolds, D. Hassabis, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538:471–476, 2016.

[16] V. Saggio, B. E. Asenbeck, P. Walther, et al. Experimental quantum speed-up in reinforcement learning agents. *Nature*, 591(7849):229–233, March 11 2021.

[17] X. Wu. Dialogue: The luckiest network on validation performs average during tests. *IEEE CDS Newsletters*, 18(1):8–11, 2024.

[18] OpenAI, J. Achiam, B. Zoph, et al. GPT-4 technical report. arXiv, 2024. arxiv.org/abs/2303.08774.

[19] J. Weng. 20 million-dollar problems for any brain models and a holistic solution: Conscious learning. In *Proc. Int'l Joint Conference on Neural Networks*, pages 1–9, Padua, Italy, July 18-23 2022. NJ: IEEE Press.

[20] M. Sur and J. L. R. Rubenstein. Patterning and plasticity of the cerebral cortex. *Science*, 310:805–810, 2005.

[21] L. B. Smith, Jones S. S, B. Landau, L. Gershkoff-Stowe, and L. Samuelson. Object name learning provides on-the-job training for attention. *Psychol. Sci.*, 13:13–19, 2002.