On the Inapplicability of Supervised Machine Learning to Evolutionary Studies

Eran Elhaik[1] and Dan Graur[2]

[1] Department of Biology, Lund University, Sölvegatan 35, Lund, Sweden, 22362
[2] Department of Biology & Biochemistry, University of Houston, Science & Research Building 2, 3455 Cullen Blvd, Suite #342, Houston, TX 77204-5001
* Please address all correspondence to Eran Elhaik at eran.elhaik@biol.lu.se

Abstract

Supervised machine learning (SML) is a powerful method for predicting a small number of well-defined output groups (e.g., potential buyers of a certain product) by taking as input a large number of known well-defined measurements (e.g., past purchases, income, ethnicity, gender, credit record, age, favorite color, favorite chewing gum). SML is predicated upon the existence of a training dataset in which the correspondence between the input and output is known to be true. SML has had enormous success in the world of commerce, and this success has prompted a few scientists to employ it in the study of molecular and genome evolution. Here, we list the properties of SML that make it an unsuitable tool in evolutionary studies. In particular, we argue that SML cannot be used in an evolutionary exploratory context for the simple reason that training datasets that are known to be *a priori* true do not exist. As a case study, we use an SML study in which it was concluded that most human genomes evolve by positive selection through soft selective sweeps (Schrider and Kern 2017). We show that in the absence of legitimate training datasets, Schrider and Kern (2017) used (1) simulations that employ many manipulatable variables and (2) a system of cherry-picking data that would put to shame most modern evangelical exegeses of the Bible. These two factors, in addition to the lack of methodological detail and the lack of either negative controls or corrections for multiple comparisons, lead us to conclude that all evolutionary inferences derived from so-called SML algorithms (e.g., discoal) should be taken with a huge shovel of salt.

Keywords: Supervised machine learning (SML), Evolutionary biology, Molecular and genome evolution, Selective sweeps

"Firstly, I never borrowed any kettle from you; second, I returned it in perfect condition; and third, it was broken when you gave it to me." Apocryphally attributed to Sigmund Freud

"Machine-learning algorithms seem to have insinuated their way into every human activity short of toenail clipping and dog washing." Benedict Carey

"The biggest issue I see with so-called artificial-intelligence experts is that they think they know more than they do." Elon Musk

Introduction

Machine learning is an analytical technique that focuses on two interrelated questions: how to construct algorithms that self-improve through experience and what kind of laws govern the learning procedure in humans and computers (Jordan and Mitchell 2015). Supervised machine learning (SML) is a useful method for predicting a small number of well-defined character states or classes using a large number of known and well-defined variables or measurements from many individuals. SML algorithms construct mathematical models based on sample data, known as the "training dataset," to make predictions or decisions on a "testing dataset" without being explicitly programmed to do so. SML algorithms aim to solve two types of problems (Libbrecht and Noble 2015): (1) the regression problem, whose target is to predict a numerical value (e.g., the average price of a house in Houston, Texas, following a major hurricane), and (2) the classification problem, whose target is a qualitative variable (e.g., determining whether or not a consumer who has ordered an item will be interested in ordering another item from a finite catalog). Here, we are only interested in the classification problem.

The immense success that SML techniques have had in the world of e-commerce (e.g., Tiwari et al. 2018; Bernardi, Mavridis, and Estevez 2019) has led some researchers to develop so-called supervised machine learning methodologies (Schrider and Kern 2016) to address evolutionary questions, particularly those aimed to clarify the relative importance of selection and random genetic drift during the evolution of genomes (Schrider and Kern 2017). Such studies have been ballyhooed as "sophisticated," "cutting-edge," "robust," and "valuable," and it has been argued that they "make a strong case for the idea that machine learning methods could be useful for addressing diverse questions in molecular evolution" (McCoy and Akey 2017).

Because SML is similar to human learning under the supervision of a teacher, we will start with a very simple illustration of SML principles as applied to a subject we are all familiar with, e-commerce. We will then outline features of SML that make it unsuitable for evolutionary studies and will end with a thorough dismemberment of Schrider and Kern's (2017) study, which concluded that soft sweeps are the dominant mode of adaptation in the human genome.

An Abecedarian Illustration of Classification by Supervised Machine Learning: Would You Like a Flashlight with Your Mineral Water?

Consider an order of mineral water from an e-commerce website that wishes to offer the most suitable complementary products. A simplistic approach to guessing what complementing products to suggest to the buyer is to consider past orders of mineral waters, which may yield answers such as a yogurt drink. Although consistent with past orders, this approach would not recognize that the order came from Houston, where a hurricane is about to hit and cause a blackout that would spoil the yogurt drink. For Houstonians, a flashlight might be the most appropriate complementing product. Whether a hurricane will indeed land, what its trajectory will be, and when it would move away are entirely irrelevant for this supervised machine learning algorithm. What matters for such an algorithm is how quickly it can identify the new correlation between the demands for bottled water and flashlights, and how quickly it can update its offerings. But what kind of flashlight can the user afford? Flashlights exhibit a wide variation in price, from $2.00 to approximately $2,100 (data from amazon.com on November 22, 2020). Here, it is useful to consider a large number of potential predictors in addition to the user's past orders. Income predictors – such as age, ethnicity, gender, credit record, education, and zip code – are extremely valuable, while other traits, such as favorite color and favorite chewing gum, are less so. If the user has neither provided personal information nor has a long history of orders, then no training data are available. However, it may still be possible to "simulate" a training dataset from the data of users from neighboring addresses, who presumably have similar income and demographic characteristics. It is possible, however, that the buyer at a posh address is only a leasee of a small room and may only earn a small fraction of the estimated mean income for that neighborhood. The website is also more likely to have a history of orders from households with a higher income and a better credit record, which may skew the correlation between the "nearest address" and income. The website might, hence, offer a much more expensive flashlight than the user can afford, and the user will take their business elsewhere.

It is obvious, therefore, that while SML can learn new correlations from the data and automatically update the offerings to increase sales when it lacks a suitable training dataset or knowledge of critical variables like demography and sex, any attempt to compensate by using proxies may turn out to be completely off base and may render the entire exercise utterly pointless. Since price variation is high, any offer based on simulations is almost guaranteed to miss the optimal price for both the buyer and the seller. In this case, simulations cannot find the best offer to maximize profit for the seller. Critically, a simulation may produce results that appear correct by selecting the best-selling flashlight ($17.99) based on previous sales (data from amazon.com on November 22, 2020). However, such results merely reinforce past patterns rather than the current state of affairs. In reality, a different flashlight, which may be cheaper or costlier, may yield higher revenues and become the best-selling flashlight.

A Case Study of Inapplicability: Schrider and Kern (2017)

In studying the role of adaptation in recent human evolution, Schrider and Kern (2017) applied an SML approach to simulated human population data that evaluated the evidence for positive selection in humans. The authors used genomic data from six human populations and a supervised machine learning procedure to classify large segments of the human genome (i.e., genetic windows) into five classes if they: (1) have experienced a hard selective sweep, (2) are linked to a hard selective sweep, (3) have experienced a soft selective sweep, (4) are linked to a soft selective sweep, or (5) have evolved neutrally, which includes regions evolving under strict neutrality as well as sequences affected by purifying selection. Their conclusion, as written in the title, was "Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome." In the paper itself, the results are presented in a more nuanced manner; while Schrider and Kern (2017) only found a minority of genomic windows to be subject to either soft or hard selective sweeps (7.6%), a large fraction of windows was found to be "linked" to completed selected sweeps either hard or soft (56.4%). Neutral evolution accounted for the evolution of only 36.0% of genomic segments.

Since there is no true training set to train the SML algorithm, Schrider and Kern (2017) simulated training datasets based on input parameters that they inferred or appropriated from the literature. In the following, we discuss their sampling transgressions, their logical missteps, their internal inconsistencies, and their corruption of statistical practice and explain why their conclusions are invalid.

Cherry-picking population data

To characterize selection in global human populations, Schrider and Kern (2017) had first to choose which populations to use. Because populations with a large variation in demographic and historical parameters due to processes such as extreme bottlenecks, a history of migrations, and complex admixture or gene flow patterns would introduce substantial uncertainties in their subsequent simulations, they decided to exclude such populations from their sample.

Schrider and Kern (2017) claimed to have selected six populations that experienced minimal influence of admixture or migration: They included GWD (Gambians in Western Divisions in The Gambia) and YRI (Yoruba in Ibadan, Nigeria) from West Africa; LWK (Luhya in Webuye, Kenya) from East Africa; JPT (Japanese in Tokyo, Japan) from Asia; CEU (Utah residents with Northern and Western European Ancestry) from Europe; and PEL (Peruvians from Lima, Peru) from the Americas. In other words, they selected three African populations and one population each from Europe, America, and Asia. Schrider and Kern (2017) motivated their choice by the homogeneity of populations over various numbers of splits (*K*) in the ADMIXTURE analysis (Alexander, Novembre, and Lange 2009) carried out by Auton et al. (The 1000 Genomes Project Consortium 2015, Extended Figure 5): "We see that for most values of *K*, each of these populations appears to correspond primarily to a single ancestral population rather than displaying multiple clusters of ancestry."

An examination of the ADMIXTURE results of Auton et al. shows that Schrider and Kern (2017) frequently violated their own criteria. For example, CEU, a population of unclear origins that shows multiple splits over various *K* values was included, while the more homogeneous British

(GBR) and Finnish (FIN) populations were excluded from the analysis. All the three African populations included in the study yield larger and more numerous splits compared to the Esans (ESN), the most homogeneous African population, which was excluded from the study. Finally, the Chinese (CDX) population was excluded despite their genetic homogeneity.

By their own admission, Schrider and Kern (2017) included the cross-continentally admixed PEL population because "among the highly admixed American samples it appears to exhibit the smallest amount of possible mixed ancestry (for most values of *K*)," so they "retained this population to have some representation from the Americas." Oddly, this "exception" clause was not applied to South Asian populations, which represent a huge part of human global genetic diversity.

The choice of populations is further questionable on several grounds. First, Schrider and Kern's (2017) inference of populations with "single ancestral population rather than displaying multiple clusters of ancestry" is doubtful since ADMIXTURE is unreliable when used to determine whether groups are pure representatives of one ancestral source and was certainly not designed to test whether two populations have mixed. Even groups that appear to be extremely homogeneous are known to have contributions from ancestrally related groups (Lawson, van Dorp, and Falush 2018). Second, applying ADMIXTURE analysis with an arbitrary number of splits yields inconsistent splits across the population panel. The appearance of "heterogeneity" is driven by populations with distinct combinations of allele frequencies rather than true homogeneity. Finally, considering the inconsistency between the selection criteria outlined by Schrider and Kern (2017) and their sample population set, it is likely that these criteria were applied *post hoc* to match populations already included in the dbPSHP database (Li et al. 2014), which the authors employed to study selection. dbPSHP has no data for the most homogeneous populations in Auton et al. (GBR, FIN, ESN, and CDX) but has data for four of the six populations analyzed by Schrider and Kern (2017): YRI, LWK, CEU, and JPT. Excepting Gujarati Indians in Houston, Texas (GIH), dbPSHP excludes all South Asians.

Studying selection in populations with complex demographic histories, extensive genetic drift, and episodic selection requires distinguishing between these processes, particularly in light of the low selection coefficient values used in the simulations (uniformly distributed between 0.005 and 0.1).

To summarize, Schrider and Kern (2017) did not follow their own criteria for selecting populations with the least complex demographic histories, which essentially reduced the accuracy of their simulations. Instead, Schrider and Kern (2017) employed a type of observational bias, called the "streetlight effect" (Freedman 2010), by only searching for preconceived answers where it is easiest to find them (i.e., in populations for which selection data were readily available).

Estimating the parameters for the demographic model

To detect sweeps, Schrider and Kern (2017) applied a maximum likelihood approach that considers simulated genomic patterns using a variety of population genetic summary statistics and classifies genomic windows as being: (1) the target of a completed hard sweep (hard), (2) closely linked to a hard sweep (hard-linked), (3) a completed soft sweep (soft), (4) closely linked to a soft sweep (soft-linked), or (5) to have evolved neutrally (neutral). Since no training dataset is available in exploratory evolutionary studies, Schrider and Kern (2017) first needed to develop a demographic model that will spew a number of summary statistics for genomic regions that have experienced simulated hard sweeps, simulated soft sweeps, or have evolved under simulated neutrality.

We emphasize that unlike in normal SML approaches, where training and testing of the classifier are two separate operations, in Schrider and Kern (2017) the classifier was never trained since no training data are available. Notwithstanding this glaring insufficiency, the authors use the terms “training” eleven times in their article. For instance, they claim that “training examples for the hard class experienced a hard sweep in the center of the central sub-window” despite the fact that the “training” data were not actual data as required for SML applications but rather simulated “data” in what we will demonstrate is an extremely problematic manner. It is, therefore, inappropriate to use terms like “machine learning” or “classifier.” Terms like “machine pretend learning” and “tarot divination” may be more suitable. In the next section, we show that even as far as building their faux “classifier” and estimating the variables on which it was based, the authors made some very unreasonable decisions.

Estimating demographic history from genomic data. Before selection can be simulated on the genomic data, Schrider and Kern (2017) had to reliably capture population-size changes over discrete time intervals for their coalescent simulation tool *discoal* to work properly (Kern and Schrider 2016). For that, they turned to the demographic model calculated by Auton et al. using the Pairwise Sequentially Markovian Coalescent (PSMC) model (The 1000 Genomes Project Consortium 2015). PSMC employs coalescent methodology to reconstruct changes in the effective population-size history over time under neutrality. Schrider and Kern extracted 26 discrete points per population from Auton et al.’s extended Figure 5 (Figure 1A), scaled them by the mean population mutation rate Θ of the population under study (which they determined as we show in the next section) and by the present-day effective population size N_0 (10,000), and included them in their simulation (the -*en* parameter) (Figure 1B). This simple procedure, which should have resulted in a similar demographic model to the one used by Auton et al., has instead resulted in inflated population sizes by a factor of up to 10^4 and a complete distortion of Auton et al.’s demographic model (Figures 1A-B). We note that PSMC’s output cannot always be reliably interpreted as plots of population-size changes, particularly if the population is admixed, in which case peaks on the demographic plot might correspond to periods of increased population structure rather than increased population size (Mather, Traves, and Ho 2020). Remarkably, Schrider and Kern (2017) noted that “these models may not accurately capture the demographic histories of the populations we examined.” However, they decided that their Soft/Hard Inference through

Classification or S/HIC method (Kern and Schrider 2016) is robust to "demographic misspecification," and hence did not expect this factor "to severely impact" their analysis.

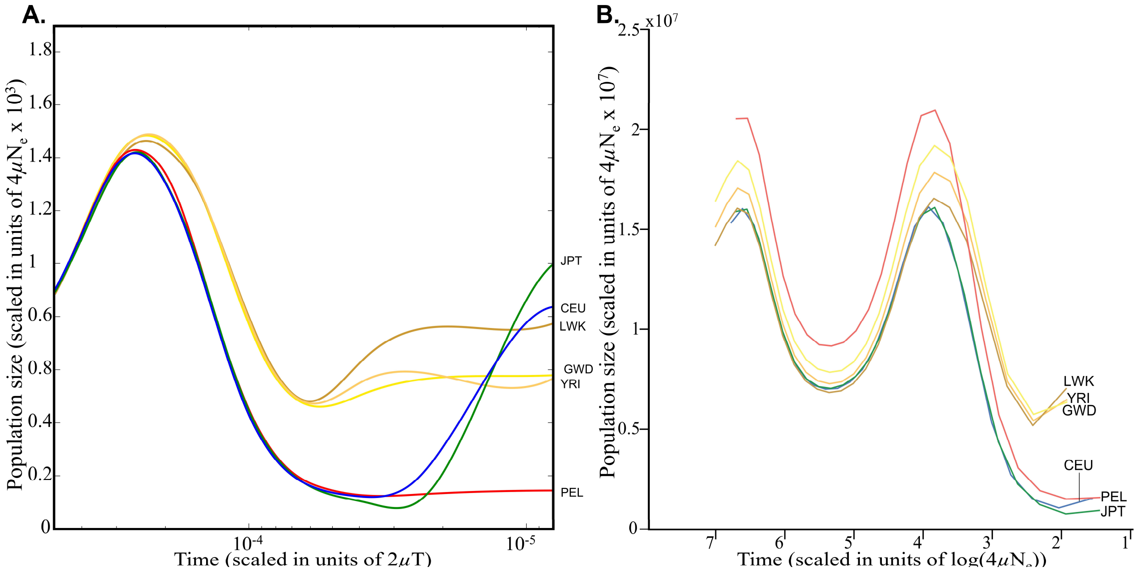

Figure 1. Population structure and demography. Changes in population size over time for six populations. **A**. Changes to the effective population sizes over time in six populations inferred by PSMC. Lines represent the within-population median PSMC estimate, smoothed by fitting a cubic spline passing through bin midpoints. The original figure was published by Auton et al. (The 1000 Genomes Project Consortium 2015, Figure 2). This figure was created using code and data provided by Dr. Adam Auton to include only the relevant populations. The plot is log-scaled for the X-axis. **B**. Schrider and Kern's (2017) recreation of Auton et al.'s figure. Schrider and Kern (2017) sampled 26 data points from (A) and scaled them by Θ and N_0. The X-axis is also Θ-scaled (to get each population on the same timescale) and log-scaled to increase the similarity with (A).

The population mutation rate (Θ). The population mutation rate (Θ) is a fundamental parameter in evolutionary biology which measures the average genomic mutation rate of the entire population or, stated differently, it describes the amount of selectively neutral diversity in a population (Milgroom 2015). Importantly, it is also necessary to estimate the effective population size (N_e). Θ is calculated as $2pN_e\mu_{tot}$, where $p = 1$ or 2 for haploids and diploids, respectively, N_e is the effective population size, and μ_{tot} is the mutation rate at the locus of interest. Schrider and Kern (2017) assumed that the present-day grid of Θ ranged from 10 to 250 and was calculated from $4N_e\mu L$. Here, μ is the mutation rate per nucleotide, which Schrider and Kern (2017) set as 1.2×10^{-8} per base pair per generation as was calculated by Kong et al. (2012) (based on Icelandic trios for all their populations), and L is the length of the segment.

Confusingly, multiple L values were used: 100,000, the only one reported in the paper; 200,000, which was used in the code (their Table S5); 2,200,000, which was used for Table S1 (D. R. Schrider, personal communication); and 1,100,000, the "correct" value for this analysis (D. R. Schrider, personal communication). Schrider and Kern (2017) neither disclosed the range of L

values nor explained them. Using the latter value of L, $N_e = 1{,}500$ for Africans and $N_e = 4{,}250$ for non-Africans from (The 1000 Genomes Project Consortium 2015), and Θ should have a theoretical range of 79 and 224, similarly to the proposed grid. Unfortunately, neither these nor the stated grid values of Θ (10–250) were used. In practice, Schrider and Kern (2017) employed Θ values that ranged from 40 to 2,200 (their Table S5) since they chose "as the final values of Θ that for which the sum of the percent deviations of the simulated from the observed means of each statistic was minimized." By inflating the range of Θ, Schrider and Kern (2017) have artificially modified the genomic mutation rate providing more opportunities for "selection" to act. As we shall next show, this choice also had the effect of bloating the effective population size.

The effective population size (N_e). Deriving N_e from the two extreme values of Θ (40 and 2,200) and the published L value ($L = 100{,}000$) yields a remarkable N_e range of $8{,}000 < N_e < 458{,}000$, i.e., values similar to those of *Plasmodium falciparum* ($210{,}000 < N_e < 300{,}000$) and *Drosophila melanogaster* (N_e=1,150,000) (Charlesworth 2009). Even using the correct L value (L=1,100,000) (D. R. Schrider, personal communication) results in an extreme range for N_e $757 < N_e < 41{,}666$, compared to the N_e calculated by Auton et al. ($1{,}500 < N_e < 4{,}250$) (The 1000 Genomes Project Consortium 2015) (Figure 1A), which Schrider and Kern (2017) relied on for their demographic model and the well-established range for humans ($1{,}000 < N_e < 10{,}000$) (e.g., Yu et al. 2004; Liu et al. 2006; Tenesa et al. 2007; McEvoy et al. 2011). This N_e is also higher than in most apes ($7 < N_e < 20{,}000$), excepting Central chimpanzee ($25{,}000 < N_e < 100{,}000$) (Marques-Bonet, Ryder, and Eichler 2009). By inflating N_e, Schrider and Kern (2017) have biased all their subsequent calculations.

Estimation of the population recombination rate (*ρ*). To understand the magnitude of the bias in Schrider and Kern's (2017) analyses, let us consider their simulation of the population recombination rate, $\rho = 4N_er$, where r is the crossover rate per base pair. The distribution of ρ was empirically found to be unimodal, with a mean of $4 \times 10^5 \times 10^{-8} = 4 \times 10^{-4}$ (Clark et al. 2003). Further, ρ is normally distributed with 80% of its values in the range of 4.5×10^{-4} to 3.5×10^{-4} (Clark et al. 2003). In contradistinction, Schrider and Kern (2017) derived ρ from an exponential distribution with a mean that is two orders of magnitude higher than the population mean (Clark et al. 2003), i.e., $8.32 \times 10^{-5} < \rho < 2\text{x}10^{-3}$ and $7.5 \times 10^{-4} < \rho < 7.5 \times 10^{-3}$, for their two N_e estimates, respectively.

Simulating genomic sequences. Deciding on the demographic model parameters for each population, Schrider and Kern (2017) have next generated genomic sequences using several approaches that we have attempted to replicate. Replication is the "cornerstone of science" and its most fundamental principle; the inability to replicate published studies has been deemed the "replicability crisis" (Baker 2016). Schrider and Kern's (2017) study is part of this crisis.

The Methods section in Schrider and Kern (2017) renders replication impossible. First, the authors employed various computational tools, but only some of them are mentioned in the paper. Second, they provided the code for only one of the tools (their Table S5). And finally, their description of the simulation is partial at times, erroneous at other times, and inconsistent with the code they have provided.

To name a few examples, Schrider and Kern (2017) wrote that "we used the program discoal (Kern and Schrider 2016) to simulate large chromosomal regions, subdivided into 11 sub-windows." However, discoal simulates very short genomic regions (in a setting of 200,000 bp, discoal simulated regions varied in size from 0–2,600 bp with a mean of 934 bp) and the subdivision into windows is part of a different package, called S/HIC (which, however, is no longer available where it was supposed to be deposited, https://github.com/kern-lab/).

In another place in the article, the authors wrote, "we simulated additional test sets of 1,000 genomic windows 1.1 Mb in length with varying arrangements of selected sites," without telling the reader which simulation tool they used.

Another typical paragraph reads: "The simulation program discoal requires some of these parameters to be scaled by the present-day effective population size; we did this by taking the mean value of θ and dividing by $4uL$, where u was set to 1.2×10^{-8} (Kong et al. 2012). The full command lines we used to generate 1.1 Mb regions (to be subdivided into 11 windows each 100 kb in length) for each population are shown in supplementary table S5, Supplementary Material online. We also simulated 1,000 test examples for each population in the same manner as for the training data." This section describes the use of at least three tools with only discoal referenced explicitly. It remains unclear how many regions were simulated since parts of the description mention 100 kb and other parts mention 1.1 Mb. Adding to the confusion is that the code (their Table S5) simulates 200 kb regions but employs parameters calculated for 1.1 Mb regions.

Bypassing training data by importing random annotation. Since Schrider and Kern (2017) lacked a true training dataset based on actual genomic data with factual annotation, they simulated their own dataset and generated their own annotation of the simulated dataset.

Innovatively, the authors devised the following quick fix to this seemingly intractable problem. They randomly selected 1.1 Mb regions from the human genome and used public datasets, such as phastCons, to annotate them. In the next stage, they generated a random sequence and copied the annotation of the real sequence onto the simulated one. For example, if the authors selected the region chr1:55000000–56100000 and if there was a conserved phastCons element at chr1:55000000–55000009, then the first ten bp of the simulated region would be marked as negatively selected (i.e., evolving in a neutral fashion). At no time were actual nucleotides from the human genome considered. In relation to our flashlight example, it is as if the seller would copy random demographic data from the better-annotated users to imaginary users. Figure 2 illustrates, in an exaggerated fashion, the ridiculousness of this method.

GCGATGACGGGTCGTAGCAATGTTCGTCTGACTATGATCTACATATTTAA

itwasthebestoftimesitwastheworstoftimesitwastheage

Figure 2. Illustration of the annotation method for the simulated "training data" used by Schrider and Kern (2017). We start with a real sequence from the human genome (top) for which an annotation exists in phastCons. Let us assume that within this sequence, one region was found to be extremely conserved (red), i.e., subject to strong purifying selection. We then take another string of letters of identical length (bottom), call it the training sequence, and annotate the corresponding positions as "evolving neutrally." If the "training" sequence is the start of the first sentence in *A Tale of Two Cities* by Charles Dickens (1859), then the string "… s the worst…" will be deemed to have evolved under purifying selection.

Classifying the simulated sequences into five classes. Schrider and Kern (2017) applied their SML classifier to the simulated data. In theory, the classifier is supposed to classify the sequences to one of the five classes (experienced a hard selective sweep, linked to a hard selective sweep, experienced a soft selective sweep, linked to a soft selective sweep, and evolved neutrally, which presumably includes regions evolving under strict neutrality as well as sequences affected by purifying selection).

The authors are very unclear about how the classifier's decisions were made. For instance, they write: "For our classifications we simply took the class that S/HIC's classifier inferred to be the most likely one, but we also used S/HIC's posterior class membership probability estimates in order to experiment with different confidence thresholds." We interpret this statement to mean that irrespective of the class inferred by the S/HIC classifier, the authors chose any threshold to make the final class determination.

An examination of Table S2 in Schrider and Kern (2017) shows that the classification is biased towards "soft selection" irrespective of the cutoff, which varies from 0.2 to 0.9 per population. In all those cases, S/HIC "classified" 72.22–99.15% of the segments as "softly selected" with a median of 93.4%. No doubt, this resulted in 73% of these sweeps to be deemed "novel" as they do not exist in dbPSHP. In order to get these newsworthy results, it was essential for the authors to preselect populations that are included in dbPSHP, regardless of their heterogeneity. The "classifications" to "hard sweep" ranged from 0.85–28.78% with a median of 6.96%. These are remarkable figures because hard sweeps are known to extremely rare in human populations (e.g., Harris, Garud, and DeGiorgio 2018). We also note that the exact balance between hard and soft sweeps depends on the yet undetermined distribution of mutational target sizes (Pritchard, Pickrell, and Coop 2010).

To summarize, classifying the genomic regions into classes required not only an SML classifier, but also a human intervention in the form of setting subjective thresholds that ranged widely in value, which generated results that are inconsistent with the literature and, hence, lack any merit. In relation to our flashlight example, it is as if the seller applied the SML classifier to newly annotated imaginary users to determine their flashlight preference from SML-recommended products but then changed this determination because of a client dataset from another seller.

There are other problems with the S/HIC classifications, one of which is that "adjacent windows are especially likely to receive the same annotation (Schrider and Kern 2016). To overcome this difficulty, a secret (i.e., unpublished) "permutation algorithm" was implemented that considered the lengths of a run of consecutive windows assigned to each class per population. The run-length distribution was then obtained from the simulated data. By the end of this procedure, the windows are thoroughly reshuffled, which addresses the concern that adjacent windows have similar annotation but not the concern that certain features and classifications are inflated due to the over-classification in the adjacent windows in the first place.

Genetic-element enrichment in selective sweeps. To complete their study that rests entirely on imaginary sequences onto which random annotations were thrust upon, Schrider and Kern (2017) ask which "biological pathways" show "a strong enrichment" in genomic regions that were deemed to have been subject to selective sweeps. Here, it was expected that the authors would test the enrichment of real biological features, such as certain metabolic pathways, certain gene families, or genes expressed in certain tissues. Alas, this is not the case. With the exception of the vague category "coding sequences," which includes open reading frames for which we have no evidence of translation, let alone function, all other features that were found to be enriched in one class or another have nothing to do with real data. For example, the transcription factor binding sites were taken from the ENCODE project (Dunham et al. 2012), which, as we all know, cannot be used to identify biological functions (Graur et al. 2013). Another so-called functional dataset is COSMIC, which is a set of somatic mutations that have been observed in cancer cells (Forbes et al. 2015). Admittedly, a minuscule minority of these mutations may play some role in tumor suppression or progression; however, the vast majority of mutations in cancer cells that are incidental have neither a causative role in cancer nor in the progression of the disease. The closest Schrider and Kern (2017) came to using real biological functions were the inferred functional categories from Gene Ontology (Gene Ontology Consortium 2015).

Here, we would like to deal briefly with Schrider and Kern (2017) statistical choices. First, in all their calculations, they used one-tailed statistical tests, which are more likely to reject null hypotheses than two-tailed tests. Second, they did not correct for multiple comparisons; hence many of their so-called significant results are very likely to be false positives. Lastly, they did not employ any type of control.

For example, Schrider and Kern (2017) reported a "dramatic enrichment" of sweeps in genes that encode proteins that interact with one another. As far as "gene networks" are concerned, however, the term "interaction" is defined very broadly. To illustrate the problematics of using interactions as a validation method, we performed the following experiment. We selected 100 random protein-coding genes that had a HUGO (Human Genome Organization) – Gene Nomenclature Committee (HGNC) symbols (Braschi et al. 2019) (https://www.genenames.org/download/statistics-and-files/) and used GeneMania (Franz et al. 2018) (https://genemania.org/) to identify genetic interactions between those genes. Of the hundred random sequences, only 17 had no "genetic interaction" with other genes. The remaining genes exhibited "genetic interactions" and "physical interactions," although none of these "interacting pairs shared a single biochemical pathway. By showing that random genes in our negative control exhibit extensive "genetic interactions," we have demonstrated that Schrider and Kern's (2017) "interacting gene networks" are a meaningless concept and that the "dramatic enrichment" is biologically insignificant .

Finally, we would like to mention a common failure to most data-driven studies—lack of any follow-up. Schrider and Kern's (2017) study contains some results that *prima facie* seem remarkable. For example, of the 19 features, the CEU population (Utah residents with Northern and Western European Ancestry) showed no significant enrichment for any annotation except "enhancers lost in humans since splitting with rhesus." In contrast, all the other populations showed significant enrichment of more than 50% of the annotated features (their Table S3). The authors offer neither an explanation for this finding nor a reason why such a result is reasonable. Similarly, ClinVar pathogenic SNPs were significantly enriched only in PEL (Peruvians from Lima, Peru) despite the fact that this population has fewer ClinVar pathogenic mutations than YRI, CEU, and CHB (Harris et al. 2018, Fig S14). As in all other cases, the authors are mum on these findings.

Discussion

While typing this manuscript, we encountered an example of the pitfalls of supervised machine learning. The misspelled word "deribed" (instead of "derived") was corrected by Microsoft Word to "derided" to yield "The authors derided ρ from an exponential distribution." This simple example shows that SML isn't a panacea for all unsolved problems.

Here, we have shown that SML is inherently inapplicable to exploratory studies on evolutionary questions related to the driving forces of evolution for the simple reason that true empirical training datasets do not exist. As far as the study by Schrider and Kern (2017) is concerned, the inappropriate methodology is further compounded by cherry-picking population data, logical missteps, internal inconsistencies, and the corruption of statistical practices. In the spirit of the motto apocryphally attributed to Sigmund Freud, we must conclude that there was no kettle, to begin with.

Despite claims to the contrary, Schrider and Kern (2017) have not provided "an excellent example of how cutting-edge methods from computer science and statistics can be successfully brought to bear on long-standing questions in evolutionary biology" (McCoy and Akey 2017). What they did provide is yet another example of the inferiority of data-driven "science" relative to hypothesis-driven science. Allen (2001) predicted that induction and data-mining, uninformed by ideas, can produce neither knowledge nor understanding. Interestingly, this prediction is a

Popperian prediction that can be disproved by a single demonstration that a hypothesis-free process, when applied to data is sufficient to produce a gain in understanding. Schrider and Kern's (2017) study constitutes no such demonstration.

Finally, we wish here to note here that the term "adaptation" in the title of Schrider and Kern (2017) may be inappropriate. The terms "adaptation" or "adaptive selection" is not a synonym for "positive selection." They should only be used when it can be shown that selection has led to a demonstrable adaptation of the organism under study to its biotic or abiotic environment.

Authors' contributions

DG initiated the study. EE carried out all the analyses. DG and EE wrote the paper. Both authors approve the paper.

Acknowledgments

EE was partially supported by the Crafoord Foundation and the Swedish Research Council (2020-03485) and the Swedish Research Council (2020-03485). The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lund, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. DG was supported by NSF DEB OPUS grant 1949939.

REFERENCES

Alexander DH, et al. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655-1664.

Allen JF. 2001. Bioinformatics and discovery: induction beckons again. Bioessays. 23:104-107.

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. Nature. 533:452-454.

Bernardi L, et al. 2019. 150 successful machine learning models: 6 lessons learned at booking.com. Pp. 1743-1751. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Braschi B, et al. 2019. Genenames. org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 47:D786-D792.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10:195-205.

Clark AG, et al. 2003. Linkage Disequilibrium and Inference of Ancestral Recombination in 538 Single-Nucleotide Polymorphism Clusters across the Human Genome. Am. J. Hum. Genet. 73:285-300.

Dunham I, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. 489:57-74.

Forbes SA, et al. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43:D805-811.

Franz M, et al. 2018. GeneMANIA update 2018. Nucleic Acids Res. 46:W60-W64.

Freedman DH. 2010. Why scientific studies are so often wrong: The streetlight effect. Discover Magazine. https://www.discovermagazine.com/the-sciences/why-scientific-studies-are-so-often-wrong-the-streetlight-effect (Last accessed Dec 3rd 2020).

Gene Ontology Consortium. 2015. Gene ontology consortium: going forward. Nucleic Acids Res. 43:D1049-D1056.

Graur D, et al. 2013. On the immortality of television sets: "function" in the Human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol. 5:578-590.

Harris AM, et al. 2018. Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. Genetics. 210:1429-1452.

Harris DN, et al. 2018. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc. Natl. Acad. Sci. U.S.A. 115:E6526-E6535.

Jordan MI, Mitchell TM. 2015. Machine learning: Trends, perspectives, and prospects. Science. 349:255-260.

Kern AD, Schrider DR. 2016. Discoal: flexible coalescent simulations with selection. Bioinformatics. 32:3839-3841.

Kong A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 488:471-475.

Lawson DJ, et al. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat. Commun. 9:3258.

Li MJ, et al. 2014. dbPSHP: a database of recent positive selection across human populations. Nucleic Acids Res. 42:D910-D916.

Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16:321-332.

Liu H, et al. 2006. A geographically explicit genetic model of worldwide human-settlement history. Am. J. Hum. Genet. 79:230-237.

Marques-Bonet T, et al. 2009. Sequencing primate genomes: what have we learned? Annu. Rev. Genomics Hum. Genet. 10:355-386.

Mather N, et al. 2020. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. Ecol. Evol. 10:579-589.

McCoy RC, Akey JM. 2017. Selection plays the hand it was dealt: evidence that human adaptation commonly targets standing genetic variation. Genome Biol. 18:1-4.

McEvoy BP, et al. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 21:821-829.

Milgroom MG. 2015. CHAPTER 4: Mutation and Random Genetic Drift. *Population Biology of Plant Pathogens: Genetics, Ecology, and Evolution*. Minnesota, US: APS Press. Pp. 59-86.

Pritchard JK, et al. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20:R208-R215.

Schrider DR, Kern AD. 2017. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Mol. Biol. Evol. 34:1863-1877.

Schrider DR, Kern AD. 2016. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. PLoS Genet. 12:e1005928.

Tenesa A, et al. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17:520-526.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature. 526:68-74.

Tiwari R, et al. 2018. Market segmentation using supervised and unsupervised learning techniques for E-commerce applications. J. Intell. Fuzzy Syst. 35:5353-5363.

Yu N, et al. 2004. Nucleotide diversity in gorillas. Genetics. 166:1375-1383.