Exploring parameter spaces with artificial intelligence and machine learning black-box optimization algorithms

Fernando Abreu de Souza,[*] Miguel Crispim Romão,[†] Nuno Filipe Castro,[‡] and Mehraveh Nikjoo[§]
LIP—Laboratório de Instrumentação e Física Experimental de Partículas, Escola de Ciências, Departamento de Física, Universidade do Minho, 4701-057 Braga, Portugal

Werner Porod[‖]
Institut für Theoretische Physik und Astrophysik, Uni Würzburg Campus Hubland Nord, Emil-Hilb-Weg 22, D-97074 Würzburg, Germany

(Received 4 July 2022; accepted 10 January 2023; published 6 February 2023)

Constraining beyond the Standard Model theories usually involves scanning highly multidimensional parameter spaces and checking observable predictions against experimental bounds and theoretical constraints. Such a task is often timely and computationally expensive, especially when the model is severely constrained and thus leading to very low random sampling efficiency. In this work we tackled this challenge using artificial intelligence and machine learning search algorithms used for black-box optimization problems. Using the constrained minimal supersymmetric standard model and the phenomenological minimal supersymmetric standard model parameter spaces, we consider both the Higgs mass and the dark matter relic density constraints to study their sampling efficiency and parameter space coverage. We find our methodology to produce orders of magnitude improvement of sampling efficiency while reasonably covering the parameter space.

DOI: 10.1103/PhysRevD.107.035004

I. INTRODUCTION

Although the Standard Model (SM) of particle physics is a hallmark of scientific achievement, it does not provide the complete picture of the fundamental degrees of freedom of the universe, leaving some phenomena unexplained. To tackle this, multiple beyond Standard Model (BSM) theories have been proposed to address a number of questions, which the Standard Model (SM) has failed to provide meaningful answers to, while successfully replicating all the features contained in the SM which have been verified experimentally. On the other hand, experiments like those at the Large Hadron Collider (LHC) at CERN are pushing the boundaries of validity of many BSM theories, while not providing so far unambiguous evidence for new phenomena beyond the SM.

In order to study the phenomenology of these BSM theories, vast parameter spaces need to be scanned to assess the values of parameters which are still valid, i.e., not in contradiction with experimental data. Such models can reach $\mathcal{O}(100)$ free parameters. However, in general, out of the virtually infinite number of possible versions of the BSM model which are represented by points scattered across the parameter space of the theory, only a tiny fraction of these points will yield predictions which are in agreement with experimental data. For instance, the minimal supersymmetric standard model (MSSM) contains 105 new free parameters, leaving a more classical examination of its parameter space rather costly and extremely time-consuming. This type of validation task can be strikingly difficult to execute, depending on the physics of the model, the number of parameters involved and the number of experimental constraints considered. This is the high-energy physics realization of a challenge known in data science as the *curse of dimensionality*, which, in this context, means that the efficiency of this exploratory analysis drops exponentially with the number of the dimensions of the parameter space.

In this regard, data-driven approaches have offered new opportunities for the investigation of high-dimensional complex problems. In recent years, artificial intelligence (AI) and machine learning (ML) have steadily become part of the tool-set of HEP researchers [1], as their algorithms provide paradigm shifting capabilities for data

[*]abreurocha@lip.pt
[†]mcromao@lip.pt
[‡]nuno.castro@fisica.uminho.pt
[§]mehraveh.n@gmail.com
[‖]porod@physik.uni-wuerzburg.de

and computationally intensive tasks. One such task is the validation of BSM theories through constraining the associated parameter space. Such task has seen recent efforts and developments of the employment of AI/ML algorithms to mitigate the burden of such scans in an attempt to increase sampling efficiency. Recent attempts at tackling this problem have deployed techniques such as deep neural networks [2,3] to try to guess if a new point is valid; Bayesian neural networks [4] to try to predict the observable value for a given parameter space point; active-learning methods [5,6] to find boundaries of valid subspaces; and generative models [7] used to resample from a collection of valid points. However, these efforts often require a large amount of data to be gathered previously for machine learning training—which presumably are hard to come by—before they can be used to suggest new points with high-efficiency, effectively not solving the sampling bottleneck.

In this work, we offer a new perspective to the sampling of new consistent model points by reframing the problem as a black-box optimization problem and bypassing the need for an initial set of sampled data. We show the efficiency of a dynamic optimization approach to the survey of two MSSM realizations, the constrained MSSM (cMSSM) and the phenomenological MSSM (pMSSM), both displaying a large reduction of the initial MSSM free parameters, by constraining the respective parameter spaces as to provide a realistic Higgs mass. For each case, we will further increase the sampling difficulty by demanding a realistic dark matter relic density.

This work is organized as follows. In Sec. II we reframed the sampling problem as a black-box optimization problem, by introducing the notion of a cost function of physical observables (themselves dependent on the parameter) that needs to be minimized. The physics cases are introduced in Sec. III, where we define the models and the observables which we will use to constrain the parameter space. Next, in Sec. IV we develop the methodology to be used for the scans, namely we introduce three AI/ML based search algorithms used for black-box optimization and how they work, as well as discussing how the scan was designed. The results of the scans and a comparison between different samplers is then discussed in Sec. V. Finally, in Sec. VI we draw the conclusions of our study and highlight the benefits and the shortcomings of the presented methodology, providing new directions of future work.

II. (RE)FRAMING THE PROBLEM

The customary approach to validate beyond the standard-model extensions against constraints and bounds on observables is to randomly sample a point, θ, from the parameter space, $\mathcal{P}$, which is then passed onto a computational routine, $\mathcal{R}$, that computes the relevant observables, $\mathcal{O}(\theta)$. The observables are then compared to experimental data, namely to check if they are within bounds (for example if the mass of an exotic new particle is above collider limits) or within uncertainties (for example if the mass of a standard model particle is within its uncertainties). If the point agrees with experimental data it is kept as a valid point, otherwise it is discarded. Depending on the difficulty of the problem at hand, i.e., how likely or not is for a random point to fit the constraints, this process can take long periods of time to collect enough valid points. On top of that, the random sampling is rather wasteful from the point of view of resources as the information of invalid points is simply discarded and not used to improve the sampling efficiency.

Previous works [2,3,8] attempted to reduce the scanning overhead by only passing to the computational routines points with a higher chance of passing the constraints. In order to achieve this, they trained machine learning models to either predict the values of the observables, $\mathcal{O}$ (using a regressor) or to predict if a point falls within experimental bounds (using a classifier). Using this methodology, they achieve a higher efficiency in the computational routine step, as only promising sampled points go through. In either case, this amounts to add a novel step in the workflow, which is the machine learning model between the sampling and the computational routine steps. Therefore, a possible difficulty with this approach is that the machine learning component might not have learned the phase space well enough to properly filter good points. Or, in other words, the efficiency of this filtering step is bounded by the amount of points sampled.

Another attempt [7], also using machine learning models, is to use generative deep learning to produce likely valid points. The authors trained normalizing flow networks on a collection of valid points in order to learn their distribution to sample more, novel points, from the same distribution. Although this approach differs from the above, as the machine learning component does not act as a filter, it faces similar obstacles as these models need vast amounts of data to be trained, for example the authors used $\mathcal{O}(10^6)$ valid points, which could be hard to collect in highly constrained scans.

In this work we present a different approach by (re)framing the problem as a black-box optimization problem to change the sampler itself. In order to shape the problem as an optimization problem, we first notice that invalid points hold a wealth a information, namely the value of the constrained observables tells us *how far* the point is from being valid. This can be captured by the constraint function, C:

$$C(\mathcal{O}) = \max(0, -\mathcal{O} + \mathcal{O}_{LB}, \mathcal{O} - \mathcal{O}_{UB}), \tag{1}$$

where $\mathcal{O}_{LB}$ ($\mathcal{O}_{UB}$) is the lower (upper) bound of the observable $\mathcal{O}$. For example, if $\mathcal{O}$ is a Standard Model mass, say the Higgs mass, $\mathcal{O}_{LB/UB} = \mathcal{O}^{\text{exp}} \mp \sigma_{\mathcal{O}}$ with $\mathcal{O}^{\text{exp}}$ the observed central value of the mass and

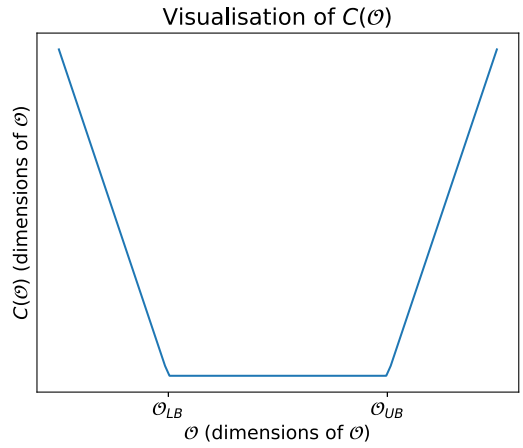


FIG. 1. Shape of the constraint function for a single observable.

$\sigma_{\mathcal{O}}$[1] the associated uncertainty. If, on the other hand, $\mathcal{O}$ is the mass of an exotic particle with experimental lower bound, O_{LB}^{exp}, then $\mathcal{O}_{LB} = O_{LB}^{\text{exp}}$ and $\mathcal{O}_{UB} = \infty$. However, we note that this function can be further expanded to included multidimensional exclusion regions, either from experiment or theory, with complicated shapes. To use those, one need to identify the *inside* region where $C = 0$ and the *outside* region, where $C > 0$ measures *how far* the point is from the interior region. In Fig. 1 we schematically show the shape of $C(\mathcal{O})$ for an observable with upper and lower bounds.

Considering now that the observables are functions of the parameters, $\mathcal{O}(\theta)$, and that the computational routines are in general black-boxes,[2] we have $C(\mathcal{O}) = C(\mathcal{O}(\theta)) = C(\theta)$. Therefore, finding valid points, θ^*, can be defined in the usual way, as the valid points (i.e., that are inside the bounds), $\mathcal{V}$,

$$\mathcal{V} = \{\theta^* : \theta \in \mathcal{P} \text{s.t.} C(\theta) = 0\}, \tag{2}$$

which can be equivalently expressed as the minimization statement

$$\mathcal{V} = \{\theta^* : \theta \in \mathcal{P} \text{s.t.} \theta^* = \operatorname{argmin} C(\theta)\}, \tag{3}$$

therefore, finding valid points for the constraints over θ amounts to minimize the function C itself and so we can treat the problem as black-box optimization problem.

[1]The notion of uncertainty depends greatly on the case study. For example, one might want to include theoretical uncertainties, which do not have a statistical interpretation, or be more lenient and allow for up to 3σ deviations from each experimental bound.

[2]There is some effort in the HEP community to produce end-to-end differentiable programming frameworks [9–11] which would allow a purely differentiable treatment of the problem. However, for BSM model building, most of the available software exists either in nondifferentiable frameworks or make use of nondifferentiable routines.

TABLE I. Physical constraints on the Higgs boson mass and dark matter relic density.

Constraint	$\mathcal{O}_{LB}$	$\mathcal{O}_{UB}$
m_{h^0}	122 GeV	128 GeV
$\Omega_{\text{DM}}h^2$	0.08	0.14

Multiple constraints can be combined using multiple C, one for each constraint. In principle, one could try to optimize against all constraints jointly as a *multiobjective optimization problem*, where one tries to find the so-called *Pareto optimal* points.[3] Here we will simplify this process and take the total constraint function as the sum of all individual constraints, as this new constraint function will still respect Eq. (2) and Eq. (3), and allows us to use single-objective optimization algorithms.[4]

In this work we will use the same constraints as in [7], namely the mass of the Higgs boson, m_{h^0}, and dark matter relic density, $\Omega_{\text{DM}}h^2$. The values of the upper and lower bounds can be seen in Table I. Both observables are known precisely from the experimental side [12]. However, their uncertainty on the theory side is signifcantly larger amounting to about $\Delta m_h \simeq 3$ GeV [13][5] and $\Delta\Omega h^2 = 0.2$ [14–16]. We will aggregate both constraints by summing the individual constraint function for each constraint

$$C(m_h \cap \Omega_{\text{DM}}h^2) = C(m_{h^0}) + C(\Omega_{\text{DM}}h^2). \tag{4}$$

The resulting function will be the loss function,

$$\text{Loss}(\theta) = C(m_{h^0}(\theta)) + C(\Omega_{\text{DM}}h^2(\theta)), \tag{5}$$

which we will minimize using black-box optimization algorithms presented in the next section.

A. Difference with fits to likelihoods

It is important to clarify the distinction between our approach and that of fitting the parameter space with likelihoods, see for example [17]. When fitting the parameter space with likelihoods, one starts with Bayes theorem

$$p(\theta|\text{data}) \propto p(\text{data}|\theta)p(\theta), \tag{6}$$

[3]In practice this means that agreement with an observable cannot be improved without simultaneously worsening at least the agreement with another one.

[4]We performed an exploratory study on different prescriptions to join multiple constraints into a single function and could not observe any difference in early results. Further exploration of this choice might yield different results and is left as future work.

[5]Strictly speaking, the theory uncertainty on m_h is smaller within the CMSSM put for the sake of comparison we assumed that this uncertainty is model independent.

where $p(\theta|\text{data})$ is the posterior of the parameters (the probability of a choice of parameters, θ, to be valid given the data), $p(\text{data}|\theta)$ the likelihood (a function which tells how likely it is the data given the choice of model and its parameters), $p(\theta)$ the parameter prior (which encodes prior distribution functions of the parameters), and we ignore the denominator which normalizes the numerator. The fit is performed by making use of Monte Carlo Markov Chain (MCMC) algorithms, which iteratively adapt the sampling region, i.e., the prior, in order to find the posterior, i.e., to tell us how likely a certain choice of parameters is given the data.

In this approach, the likelihood functions can be constructed from some observational data, for example a Gaussian where the mean and the standard deviations are provided by an observation, or can be provided by the experiments themselves (see the efforts of some collaborations to provide likelihoods and other experimental data-derived statistical functions [18]), over which the MCMC algorithm continuously samples and evaluates the priors in order to find the posterior distribution in a slow and computationally expensive process.[6] At the end of this process, a collection of points—each retained according to its probability of being valid—is obtained and from which a posterior distribution can be represented via histograms or other density estimators, with longer Markovian chains producing a better description of the posterior. MCMC fits discard many points, as they are only kept up to a probability of being valid, and can struggle to converge in high-dimensional spaces.

In our approach we are not looking for the posterior of the parameters given the data. This means that we are not concerned about how likely a point is given the data, i.e., the resulting distributions we obtain should not be interpreted as posteriors. We are concerned about how quickly and efficiently we can find regions and points of the parameter space which are valid. This means that we have to define what bounds on observables we are willing to accept, cf. Table I, and we do not have to concern ourselves with the explicit form of the likelihood. Indeed, the fact that we do not need a likelihood has its advantages, as our approach allows us to use bounds on masses or couplings of exotic physics by adding the appropriate constraint function, C, to better guide the sampler, whereas such information cannot be used in fits with likelihood functions.

[6]It is known that MCMC algorithms struggle with the so called *curse of dimensionality*, i.e., with highly dimensional priors. There has been a considerable effort to mitigate this by using neural network approximators, which have been already used to perform these fits [19]. The usage of neural networks has the added advantage of that they are differentiable and therefore easy to incorporate in MCMC algorithms that make use of derivatives, such as the Hamiltonian variation.

III. THE PHYSICS MODELS CONSIDERED

We take here the MSSM as an underlying test model. On the one hand, it remains an appealing SM extension, as it provides solutions to the most prominent shortcomings of the latter. In addition to solving the hierarchy problem related to the mass of the Higgs boson [20,21], the model includes a viable candidate for the observed cold dark matter (CDM) in the Universe, namely the lightest of the four neutralinos. On the other hand it can be formulated either as a high scale model, where only a few parameters are given, for example at the scale of grand unification. A prominent example is the constrained MSSM (cMSSM) [22], which is defined in terms of four parameters and the choice of a particular sign (phase). It can equally well be formulated as a low scale theory taking the soft SUSY breaking parameters freely at the electroweak scale. A popular variant is the so-called pMSSM [23] which takes into account the most stringent constraints from low energy data by setting flavor mixing entries to zero and neglecting possible complex phases.

Supersymmetric models are characterized via the superpotential and the soft SUSY breaking Lagrangian. The superpotential of the MSSM is given as

$$W_{\text{MSSM}} = -\varepsilon_{ab}\mu\hat{H}_1^a\hat{H}_2^b + \varepsilon_{ab}(\hat{H}_1^a\hat{L}^b Y_e\hat{E}^c + \hat{H}_1^a\hat{Q}^b Y_d\hat{D}^c + \hat{H}_2^b\hat{Q}^a Y_u\hat{U}^c) \tag{7}$$

ε is the totally antisymmetric SU(2) tensor, Y_i are the Yukawa couplings and μ is the Higgs/Higgsino mass parameter. The superfield $\hat{F}$ $(F=H_d,H_u,Q,L,D^c,U^c,E^c)$ contains the fermionic and bosonic degree of the field F. Here we have only included terms conserving R-parity. The soft Lagrangian is parametrized as

$$\begin{aligned}\mathcal{L}_{\text{soft}}^{\text{MSSM}} = &-\frac{1}{2}(M_1\tilde{B}\tilde{B}+M_2\tilde{W}\tilde{W}+M_3\tilde{g}\tilde{g}+\text{H.c.})\\ &- m_Q^2\tilde{Q}^\dagger\tilde{Q} - m_L^2\tilde{L}^\dagger\tilde{L} - m_u^2\tilde{U}^*\tilde{U} - m_d^2\tilde{D}^*\tilde{D} - m_e^2\tilde{E}^*\tilde{E}\\ &-(T_U\tilde{U}^*H_u\tilde{Q}+T_D\tilde{D}^*H_d\tilde{Q}+T_E\tilde{E}^*H_d\tilde{L}+\text{H.c.})\\ &- m_{H_u}^2H_u^*H_u - m_{H_d}^2H_d^*H_d - (bH_uH_d+\text{H.c.}).\end{aligned} \tag{8}$$

where $\tilde{\phi}$ denotes the superpartner of a generic SM particle ϕ. We neglect in the following all phases and flavor mixing entries. In this approximation one can write the trilinear parameters T_i as $T_i = A_iY_i$. One has in total even in this simplified version 31 unknown parameters. Two of the four parameters in the Higgs sector $(\mu, b, m_{H_u}^2, m_{H_d}^2)$ are traded for M_Z^2 and $\tan\beta = v_u/v_d$ where $v_{u,d}$ are the vacuum expectation values of the Higgs bosons. In this way one ensures automatically that one complies with the precise measurement of the Z-boson mass and that one is in a minimum of the potential where $\text{SU}(2)_L \times \text{U}(1)_Y$ is correctly broken.[7]

[7]However, this does not necessarily imply that this is the global minimum of the potential, see e.g., [24,25] and references therein.

TABLE II. Parameters and their bounds of the pMSSM model.

Parameter	Values	Description	SPHENO input code
m_0	[0, 10] TeV	Soft scalar mass	MINPAR: 1
$m_{1/2}$	[0, 10] TeV	Soft fermion mass	MINPAR: 2
A_0	$[-6m_0, 6m_0]$	Trilinear soft coupling	MINPAR: 5
$\tan\beta$	[1.5, 50]	Tan beta	EXTPAR: 25

In the following we will focus on the Higgs mass and the dark matter relic density as observables as already mentioned above. We summarize here a few key aspects of these observables as this will be helpful to understand some aspects of our findings. In contrast to the SM, the mass of the Higgs boson is not an independent quantity in supersymmetric models. Within the MSSM it is bounded from above by M_Z at tree level and large loop corrections are needed to bring it to the observed value of about 125 GeV. The required large coupling is given by the top Yukawa coupling and consequently the largest contribution is given by loops containing top quarks or stops, see [13] for a recent review. The relative large value of the Higgs mass m_h implies that one needs either rather heavy stops and/or a large left-right mixing in the stop sector. The mixing is controlled by the parameter A_t. The observed relic density can be explained by the lightest neutralino which is stable if it is the lightest supersymmetric particle (LSP) and if R-parity is conserved. Its dark matter properties depend strongly on its nature, see, e.g., [26] for a recent review, which in turn depends on the hierarchy of the parameters M_1, M_2, and μ. Besides its nature, which determines the annihilation rates into SM particles, the relic density will also depend on the nature of the next to lightest supersymmetric particle(s) as this might open coannihilation channels if the mass difference is not too large [27]. Moreover, there is also the possibility of an s-channel resonance via the pseudoscalar Higgs boson if the mass of this Higgs boson is about twice the mass of the neutralino [27].

We will use SPHENO [28,29] for the calculation of the masses and mixing angles which serves as input for MICROMEGAS [30,31] which calculates the relic dark matter density. The data transfer between these programs is handled using the SLHA format [32,33]. In SPHENO the MSSM is matched onto the SM at the scale $M_{\rm SUSY} = \sqrt{m_{\tilde{t}_1} m_{\tilde{t}_2}}$ [34] where $m_{\tilde{t}_i}$ are the masses of the two stops. In this way one ensures a proper decoupling of the SUSY particles if their masses get very large compared to the electroweak scale.

A. cMSSM

The cMSSM is defined in terms of four parameters: at the scale of grand unification (GUT scale) one provides a common scalar mass parameter m_0 for the sfermions and Higgs bosons, a common trilinear coupling A_0 between sfermions and Higgs bosons as well as a common gaugino mass parameter $m_{1/2}$. In addition one fixes $\tan\beta = v_u/v_d$ at the electroweak scale. The modulus of the superpotential parameter μ is fixed by the requirement of getting the correct value for M_Z but its sign or more generally its phase is still a free parameter. We assume for this part of the investigation $\mu > 0$. We give in Table II the ranges of the parameters considered as well as the corresponding entry within the SLHA format for the convenience of the reader.

The overall mass scale of the stops is roughly given by $\sqrt{m_0^2 + 4m_{1/2}^2}$ and the left right mixing parameter $A_t \simeq -2m_{1/2} + 0.2A_0$ in case of small $\tan\beta$. Approximate formulas for these parameters valid also for large $\tan\beta$ can be found in [35]. Thus, one needs in general sizeable values of m_0 and $m_{1/2}$ to explain the observed Higgs mass [13,36].

The required value of the DM relic density can only be achieved in particular slices of parameter space where one has either coannihilation or a Higgs-funnel resonance if the LSP is binolike [37,38]. The coannihilation usually requires a light stau or a light stop within the cMSSM [39,40]. A winolike LSP is not possible in this model but there is a slice where the LSP is Higgsino-like [37,38].

B. pMSSM

In this model one defines the parameters at the scale $M_{\rm SUSY}$ neglecting all CP phases and flavor mixing parameters. In addition one assumes that the mass parameters of the first two generations sfermions are equal for particles with the same quantum numbers. Moreover, the A-parameters of the first two generations are set to zero. This amounts in 19 free parameters which are summarized in Table III where we give again the corresponding entries for the SLHA convention in the last column. The ranges for the parameters are chosen such that existing LHC bounds on the various supersymmetric particles are taken into account automatically. For certain combinations those bounds could be lowered but we do not expect that these additional points give additional features for the observables considered.

This additional freedom decouples completely the dependence of the two observables pMSSM on the parameters. The stop mass parameters are still the most important ones for the Higgs mass. However, for the relic density several additional possibilities open up. First, also the neutral wino becomes an accessible dark matter candidate. Second, in this class of models one can adjust the

TABLE III. Parameters and their bounds of the pMSSM.

Parameter	Values	Description	SPheno input code
$\lvert M_1\rvert$	[0.05, 4] TeV	Gaugino (Bino) mass	EXTPAR: 1
$\lvert M_2\rvert$	[0.4, 4] TeV	Gaugino (Wino) mass	EXTPAR: 2
M_3	[1, 4] TeV	Gaugino (gluino) mass	EXTPAR: 3
$\lvert\mu\rvert$	[0.4, 4] TeV	Bilinear Higgs mass	EXTPAR: 23
$\lvert A_t\rvert$	[0, 6] TeV	Top trilinear coupling	EXTPAR: 11
$\lvert A_b\rvert$	[0, 4] TeV	Bottom trilinear coupling	EXTPAR: 12
$\lvert A_\tau\rvert$	[0, 4] TeV	Tau trilinear coupling	EXTPAR: 13
m_A	[0.1, 4] TeV	Pseudoscalar Higgs mass	EXTPAR: 26
$\tan\beta$	[1, 60]		EXTPAR: 25
m_{L_1}	[0.1, 4] TeV	1st generation left-handed slepton mass	EXTPAR: 31
m_{e_1}	[0.1, 4] TeV	1st generation right-handed slepton mass	EXTPAR: 34
m_{L_2}	m_{L_1}	2nd generation left-handed slepton mass	EXTPAR: 32
m_{e_2}	m_{e_1}	2nd generation right-handed slepton mass	EXTPAR: 35
m_{L_3}	[0.1, 4] TeV	3rd generation left-handed slepton mass	EXTPAR: 33
m_{e_3}	[0.1, 4] TeV	3rd generation right-handed slepton mass	EXTPAR: 36
m_{Q_1}	[0.7, 4] TeV	1st generation left-handed squark mass	EXTPAR: 41
m_{u_1}	[0.7, 4] TeV	1st generation right-handed u-type mass	EXTPAR: 44
m_{d_1}	[0.7, 4] TeV	1st generation right-handed d-type mass	EXTPAR: 47
m_{Q_2}	m_{Q_1}	2nd generation left-handed squark mass	EXTPAR: 42
m_{u_2}	m_{u_1}	2nd generation right-handed u-type mass	EXTPAR: 45
m_{d_2}	m_{d_1}	2nd generation right-handed d-type mass	EXTPAR: 48
m_{Q_3}	[0.7, 4] TeV	3rd generation left-handed squark mass	EXTPAR: 43
m_{u_3}	[0.7, 4] TeV	3rd generation right-handed u-type mass	EXTPAR: 46
m_{d_3}	[0.7, 4] TeV	3rd generation right-handed d-type mass	EXTPAR: 49

parameters such, that all electroweakly interacting supersymmetric partners can in principle be close in mass to allow for coannihilation. This is even true for squarks because the required small mass difference leads to very soft jets at the LHC which drastically reduces the bounds from direct searches [41–43]. In particular light sleptons of the first generations can be light covering a part of the parameter space where the observed deviation of the anomalous magnetic moment of the muon can be explained [44].

IV. SAMPLERS AND METHODOLOGY

Having reframed the parameter space scan as an optimization problem, and the physics cases that we will use in this work, we now present the samplers and the HEP computational routines that we will use.

The three sampling algorithms presented here, in addition to the random sampler that we will use as a baseline to compare their behavior, operate in different ways and are representative of big classes of black-box optimizers. The purpose of using these three is to evaluate and assess how different approaches to black-box optimization can impact the final result in terms of both sampling efficiency, i.e., how easily they produce valid points, and coverage of the parameter space, i.e., how much of the parameter space was explored and if the samplers are focusing on subsets of it. Indeed, these two characteristics present two opposing forces, which in machine learning and artificial intelligence literature is commonly known as *exploration-exploitation trade-off*, where the former accounts for the capacity to explore the breadth of the parameter space, whereas the latter accounts for the inclination of an algorithm to exploit the information to get to a minimum (which could be local) as fast as possible.

As the approach presented herein is agnostic of the physics case being studied, and considers the HEP computational routine to be a black-box function, it is also important to point out that all algorithms used in this work are gradient-free, i.e., they do not rely on any gradient computation of the loss function. This is important as our loss function is a black-box function produced by the HEP routine which generally cannot be differentiated. In principle, one could compute numerical derivatives by evaluating in the infinitesimal neighborhood of a point, however this would lead to too many black-box routine evaluations and to slower sampling speeds. Alternatively, one could produce a transparent box routine through which derivatives could be computed. Such approach, usually referred as *differential programming*, would allow for different approaches making use of autodifferentiation such as those usually used in neural networks training. Unfortunately, this represents a change of paradigm in routine development, which is not yet customary in HEP and therefore outside the reach of this work.

A. Tree-structured Parzen estimator

The tree-structured Parzen estimator (TPE) [45–47] is a Bayesian optimization algorithm. Such algorithms are composed of primarily two components: a surrogate model and an acquisition function. The surrogate model is a probabilistic model which iteratively approximates, i.e., learns, the cost function produced by the black-box, i.e., it approximates $p(\mathrm{Loss}(\theta)|\theta)$. The acquisition function is a prescription to choose which point, as sampled using the information gathered by the surrogate model, is used to evaluate the black-box in the subsequent iteration.

Due to the probabilistic nature of the surrogate model, Bayesian optimization algorithms have a natural predisposition to explore the parameter space early on, when few points have been sampled and the uncertainty about the cost function is high. As more points are used to learn the cost function, the acquisition function tends to prefer better points more confidently, moving the algorithm to an exploitation phase.

Each Bayesian optimization algorithm has its own design for the surrogate model and acquisition function. The TPE uses Bayes theorem starting from the surrogate model

$$p(\mathrm{Loss}(\theta)|\theta) = \frac{p(\theta|\mathrm{Loss}(\theta))p(\mathrm{Loss}(\theta))}{p(\theta)} \tag{9}$$

which is simplified by separating the points into two densities, one for good points, $g(\theta)$, and another for bad points, $l(\theta)$,

$$p(\theta|\mathrm{Loss}(\theta)) = \begin{cases} l(\theta), & \text{if } \mathrm{Loss}(\theta) \geq \mathrm{Loss}^* \\ g(\theta), & \text{if } \mathrm{Loss}(\theta) < \mathrm{Loss}^* \end{cases}, \tag{10}$$

where Loss^* is a cutoff value which splits points into good and bad.[8] The distinction between good and bad is made through a quantitative heuristics built-in routine, see [45] for details,[9] and the densities $g(\theta)$ and $l(\theta)$ are approximated using Gaussian mixture models. The crucial intuition is that sampling is performed on the good point distribution, $\theta' \sim g(\theta)$, and the quality of a new sampled point, θ', is a function of the likelihood ratio between both densities, $g(\theta')/l(\theta')$. Points which have a high likelihood ratio between both densities are kept, given to the black box, and the process repeats until a limit of trials has been performed. Early on, both distributions will be similar and diffuse, leading to a high exploration of the space. As more points allow for a better distinction between good and bad points, TPE will start to favor exploitation of the good points distribution. However, since each sampling step is stochastic, and the decision to retain or not a point is made by comparing likelihoods of two density approximations, TPE will always retain a certain level of exploration, which in principle might lead to a better coverage of the parameter space.

It is important to note that the value of the loss, $\mathrm{Loss}(\theta)$, is only used to separate points using a heuristic cutoff value, i.e., TPE does not learn $p(\theta|\mathrm{Loss}(\theta))$ as it happens with other Bayesian optimization algorithms. In other words, the value of the loss is only used to *sort* the points, an operation which is independent of the nominal order of magnitude of the value of the loss function.

B. Nondominated sorting genetic algorithm II

Nondominated sorting genetic algorithm II (NSGA-II) [48] is a genetic evolutionary algorithm. Genetic algorithms are characterized by a loop where a subset of points, a population, is improved by selection. This loop is called a generation and has four main steps

(1) Evaluation: Where we compute the fitness function, in our case the loss function, for all members of the population.
(2) Selection: Where the points are sorted and the best ones are selected to breed and generate a new generation.
(3) Recombination: Where pairs of parents are combined for mating and an offspring is generated by mixing the genes of the parents.
(4) Mutation: Where some elements of the offspring see their genes randomly changes.

These steps are repeated until a stopping criteria is met, for example a maximum number of generations. In our implementation, the genes of each individual are the values of the parameters. Evaluation is carried out by passing the parameter space point through the black-box and testing the produced observables with the loss function. In each generation, the members are ranked by the value of the respective loss. A new generation is produced by keeping the best elements, the *elite*, and new elements are produced through *offspring*, where genes are exchanged between two parents via *cross-over* to produce a new member, exploiting the features of the elite parents. When new members are generated, *mutations* can be applied to some genes (i.e., values of some of the parameters) randomly to increase exploration by applying Gaussian noise to the values of the parameters. As with any genetic algorithm, NSGA-II uses $\mathrm{Loss}(\theta)$ to sort the members of the population to select the *elite* that will produce the offspring.

Genetic algorithms start off with a randomly initialized population and begin exploiting the best elements of the population after a single generation. As the number of generations increase, the population becomes more and

[8]Notice that in our case the black-box is deterministic, i.e., $p(\mathrm{Loss}(\theta))$ is 1 if the point has produced physical observables, and 0 if it is not physical, i.e., if SPHENO does not produce a valid spectrum. This also includes the cases where the LSP is charged.

[9]The prescription to define Loss^* is akin to a rolling quantile which becomes progressively smaller as the number of iterations grows.

more specialized and its members similar among themselves, hindering exploration. The mutation step can produce some exploration later on, but too much mutation will prevent convergence. Furthermore, because new points are obtained by mixing the values of previous points, genetic algorithms are especially suitable for combinatorial and constraint satisfaction problems, this is because some of such problems can be solved by finding good combinations of parameters. This means that genetic algorithms tend to produce characteristics that survive multiple generations, which are called schemas, producing clustered values of the parameters.

In NSGA-II the members of the population are first sorted into groups regarding their loss function performance, and then further sorted by crowding distance to mitigate the risk of getting the population stuck in a local minima. NSGA-II is specially crafted for multiobjective optimization problems. For single objective, as we perform here, it resembles a traditional genetic algorithm. The study of its performance and behavior for multiobjective problems is left for a future work.

C. The covariance matrix adaptation evolution strategy

The covariance matrix adaptation evolution strategy (CMA-ES) [49] belongs to the class of evolutionary strategy algorithms that do not implement genetic encoding to produce offspring. In comparison to the genetic algorithms presented in the previous section, evolutionary strategy algorithms do not have parents producing offspring by interchanging genes. Instead, they use the best members of the population to approximate a localized density from which they sample the new generation. In order words, a new generation is produced from the statistics learned from the previous generation.

The CMA-ES algorithm samples new candidate points from a multivariate normal distribution, for which the mean—that controls the direction of the evolution—and the covariant matrix—which captures the relations between parameters—are adapted, i.e., learned, from the previous points. This is the sense where this is an evolutionary algorithm, as new points are produced through the information of the previous ones, but there is no direct parent to offspring genetic crossover, instead the new members of the population are derived from moving statistics.

The mean of the distribution is updated as to maximize the likelihood under the multivariate normal distribution of the best performing points. More specifically, the mean vector of the multivariate normal is updated through a rolling mean with the best points (usually half of the population). CMA-ES is expected to converge rapidly, as the (approximate) covariant matrix works as a proxy for the second derivative of the loss function, i.e., the Hessian, resembling a higher-order optimization process. In this sense, CMA-ES is very similar to gradient descent algorithms, where a point is iteratively moved along the opposite direction of the gradient of the loss function. However, gradient descent algorithms require not only evaluating the value of the loss for a point, they also require computing its derivative, which can be computationally heavy. Instead, CMA-ES use a population to approximate this descent, leveraging information which replaces the Hessian for a fast convergence. Intuitively, CMA-ES can be thought as of a herd of animals descending from the mountains, meeting in the valley, and moving together to the plane. Therefore, one expects CMA-ES to produce points very close to each other as it quickly converges to a minimum of the loss function.

Although it uses a multivariate normal, CMA-ES is fundamentally different to TPE. In TPE a Gaussian mixture model is used to approximate point density, from which new points are sampled. Gaussian mixture models can fit multimodal distributions, and provide a rich description of point density. On the other hand, a single multivariate normal, as used in CMA-ES, can only describe a single mode from which new points are then suggested. In particular, CMA-ES will focus on valid points around the current best mean, whereas TPE can maintain information of all previously tried points. Therefore, we expect CMA-ES to be the most eager algorithm of the three, although its reliance of a single multivariate normal might prevent it from achieving fast convergence in highly multidimensional spaces due to the so-called *curse of dimensionality.*

D. Implementation details

We have introduced three different black-box optimization algorithms that cover three distinct classes: a Bayesian optimization algorithm, a genetic algorithm as well as an evolutionary algorithm. This will allow us to explore the differences and nuances of each algorithm when applied to our problem. We now describe how our experiment was conducted.

For the numerical routines to compute physical observables, we have used SPHENO-4.0.5 [28] and MICROMEGAS_5.2.13 [31], in order to calculate the Higgs mass and dark matter relic density, respectively. We compute the mass spectrum using SPHENO GUT scale input parameters for cMSSM (cf. Table II), and SUSY scale for the pMSSM (cf. Table III). SPHENO output spectrum files are used as inputs of `micrOMEGAs` to calculate the dark matter relic density. We performed two parallel studies, with and without dark matter relic density constraint, while keeping the Higgs mass constraint for both of the studies.[10] We discard and penalize unphysical points involving charge-breaking vacua from charged scalars and charged LSP.

[10]The physics choice was made as to have a similar study to [7]. However, their implementation relies on SOFTSUSY version 4.1.0, whose routines to compute the parameters relevant to the Higgs mass differ, leading to lower sampling efficiencies. Nonetheless, we decided to keep these physics cases.

This is done by assigning to these points an infinite value to their loss. Since all algorithms use the loss values to sort candidate points, this guarantees that unphysical points will become less and less likely to be suggested as good points.

The parameter spaces have been sampled and the loss optimized using `Optuna_2.8.0` [50], with the built-in Random, TPE, NSGA-II, and CMA-ES samplers. We changed the default settings for the TPE sampler to `multivariate=True`, in order for the Gaussian mixtures to learn the correlations between the variables. The heuristic to calculate Loss* was left as the default, which is defined as a gamma function, γ, that cuts off the trials at number

$$\gamma(n) = \min(\mathrm{ceil}(0.1 \times n), 25), \tag{11}$$

where n is the trial number, ceil the ceiling operation which rounds up its argument. This function returns an integer, which sets the number of best trials to be used to compute the good point distribution, $g(\theta)$. In this default variation, we see that $g(\theta)$ will be at most approximated by the best 25 points. The NSGA-II parameters were set to default, which means that the cross-over probability, i.e., the probability of a pair of parents to produce offspring, was set to 0.9, and the mutation probability for each gene is 1/number of parameters, meaning that each element of the offspring has, on average, one parameter mutated. For the CMA-ES sampler to `restart_strategy='ipop'`, which is a heuristic to restart the multivariate normal if convergence is seemingly stuck in a local minimum, as to force exploring new regions.

We did not sample directly from the parameter space definitions in Tables II and III. Instead, we sampled from a hyper-cube of size 1, which we call the box parameter space, $\hat{\mathcal{P}}$, which has the same dimension as the physical parameter space, $\mathcal{P}$. A box parameter space point, $\hat{\theta} \in \hat{\mathcal{P}}$, is then reshaped to be in $\mathcal{P}$ before being fed to the computational routine.[11] This allows us to treat all the parameters as ranging the same nominal values, in this case between 0 and 1, to better derive comparing metrics, discussed bellow. We notice that the map is isomorphic, so a point $\hat{\theta}$ in $\hat{P}$ maps to only one point in $\mathcal{P}$ and vice-versa, so they can be thought as the same.[12]

[11]These transformations are mostly linear transformations to recenter and resize the interval from [0, 1] to the intended range. The exception being the parameters sampled from two disjoint intervals. Take for example the μ in the cMSSM case has values over $[-4, -0.4] \cup [0.4, 4]$ TeV. We first sample from $\hat{\mu} \sim [0.1]$, then reshape it to include negative numbers $\hat{\mu} = 2 \times (\hat{\mu} - 0.5)$, then we keep its sign aside, and rescale and recenter its value to match the desired interval $\mu = \mathrm{sign}(\mu')(|\mu'| \times 3600 + 400)$. This way we avoid having to perform a separate sampling for the sign and all parameters are sampled from [0, 1].

[12]In genetic algorithm terminology $\hat{\theta} \in \hat{\mathcal{P}}$ is the genotype representation of the point/individual and $\theta \in \mathcal{P}$ is the phenotype representation of the point/individual.

In early exploratory runs, we observed that the convergence speed for the TPE became progressively slower as the number of successive trials reached a few thousands. This is understood as the surrogate model in TPE, a Guassian mixture model, is known to have a high computational complexity, which makes it forbiddingly slow for long runs. In order to mitigate this, each scan for each sampler was limited to 2000 sequential steps, called trials, and repeated 500 time, which we call episodes, totaling one million points for each combination.

E. Evaluating the samplers

In order to compare the samplers, we developed three different metrics. The first one, efficiency, is just the percentage of valid points found by the sampler

$$\mathrm{Efficiency} = \frac{\#\,\mathrm{valid\ trials}}{\#\,\mathrm{total\ trials}}. \tag{12}$$

This is the most intuitive metric to compare samplers, as we want highly efficient samplers to tackle difficult constraints. However, we need to have a measurement on how the sampler is exploring the parameter space. We need a quantitative way of measuring how much of the parameter space each sampler has explored. To do this, we introduce two metrics.

The first metric to measure the width of the exploration is the mean Euclidean distance between the sampled valid points. A sampler that explores narrow regions of the parameter space is expected to produce smaller mean distances between sampled valid points, whereas an exploration oriented sampler will produce high mean distances:

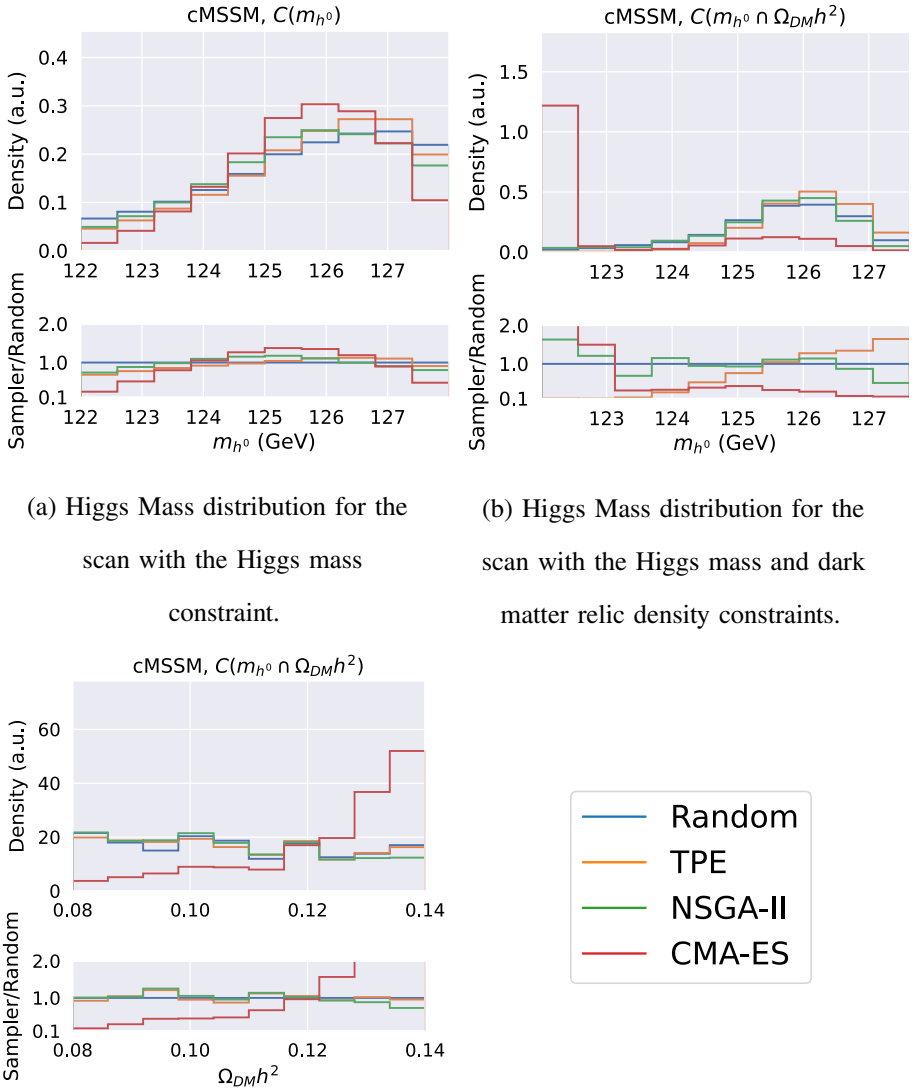
$$\mathrm{Mean\ Euclidean\ Distance} = \mathbb{E}_{\hat{\theta}_i, \hat{\theta}_j \in \mathcal{V}}\left[\sqrt{(\hat{\theta}_i - \hat{\theta}_j)^2}\right], \tag{13}$$

where $\hat{\theta}_i$, $\hat{\theta}_j$ are any two points in the valid region of the parameter space, $\mathcal{V}$, as seen in the box parameter space, $\hat{P}$. The reason why this metric is obtained in the box parameter space is that higher nominal values would dominate the value of the distance, and dilute the impact of sparser distributions in smaller valued parameters. For a hypercube of dimension d and size 1, the maximal distance between two points is given by the longest diagonal, $\sqrt{d}$, and it serves as gauge to the size of the box parameter space.

The second metric to measure the exploration is the Wasserstein distance (WD). Given two univariate distributions, $f(u)$ and $g(u)$ over the same domain, $u \in U$, and their cumulative distribution functions, $F(u)$ and $G(u)$, the Wasserstein distance between the two distributions is

$$WD(f, g) = \int_U |F(u) - G(u)| du, \tag{14}$$

and measures how different the two distributions are. We will use this to measure how much of the parameter space is



(a) Higgs Mass distribution for the scan with the Higgs mass constraint.

(b) Higgs Mass distribution for the scan with the Higgs mass and dark matter relic density constraints.

(c) Dark matter relic density distribution for the scan with the Higgs mass and dark matter relic density constraints.

FIG. 2. Top panels: target observables distributions for the cMSSM scans. The resulting valid points histograms for each sampler are produced from joining all the episodes. Bottom panels: the ratio between the histogram of the random sampler with the remaining samplers. In all cases the histograms represent a density, which the area equals to one.

being covered by different samplers. To this effect, we compute WD for each parameter distribution of valid points against the uniform distribution, which the cumulative distribution function is just the straight line starting at the origin and ending at $(\max(u), 1)$.[13] Since the uniform distribution over a parameter is the maximal coverage possible in that dimension of the parameter space, this quantity measures how far off a distribution of valid points is from covering all possible values.

We notice however, as it was highlighted in Sec. II A, our goal is not to fit the posterior distribution of the points, therefore this metric should not be taken as a dissimilarity measurement between the obtained distributions here and distributions obtained through a fit with likelihoods.

[13]In fact, we computed over the distributions of the parameter values in the box space to simplify the process, where the endpoint is (1,1).

Instead, this metric is a proxy to how far a sampler is from exploring the whole parameter space. We also note that the distributions of the random sampler are not expected to have vanishing WD with the uniform distribution, as the random sampler parameter distributions of valid points are distorted by the constraints and therefore will not be uniform distributions themselves.

The pairwise Euclidean distances were computed using `numpy` [51] `pdist` functions. The cumulative distribution functions of the parameters were computed using `statsmodels` [52] `ECDF` class. The Wasserstein distance was computed using the `SciPy` [53] `wasserstein_distance` function. Data manipulation was done with `pandas` [54], and for data visualization we used `matplotlib` [55], `seaborn` [56], and `mplhep` [57].

V. RESULTS

We now present the results of the scans produced with the different samplers for the different Physics cases. For both the cMSSM and the pMSSM as introduced in Sec. III, we performed two scans: one with the Higgs mass constraint only, and another with both the Higgs mass and the dark matter relic density constraints, with bounds defined in Sec. II.

A. Target observables and sampled parameters

In this section we present the distributions of the target observables and scatter plots of some of the parameters.

1. cMSSM

In Fig. 2 we can see the distributions for the Higgs mass and the dark matter relic density for the cMSSM for each sampler, in the top panels. In the bottom panels we show the ratio of the histogram of each sampler against the random sampler to further illustrate how different samplers produce different distributions.

We notice that TPE and NSGA-II both produce distributions relatively close to the random sampler ones, while CMA-ES exhibits more pronounced deformations. In more detail, we see how CMA-ES seems to center the distributions far closer to the edges of the allowed values for the case where we include the dark matter constraint. This region is characterized by stau coannihilation (see also discussion of Fig. 6) and in the corresponding region of parameter space the two constraints compete against each other as the Higgs mass tends to be on the lower side whereas the relic density tends to be on the larger side. The observed feature might be due to the way that CMA-ES works, akin to a gradient descent, looking for a path to minimize the loss which might force it to look for a path of least resistance in the parameter space.

In Fig. 3 we present the average over episodes of the Wasserstein distance for each distribution. This measures how much the distributions of valid points differ to the uniform distribution, as to quantify the parameter space coverage of each sampler. Smaller (larger) values of the Wasserstein distance mean that the distribution is more similar (different) to a uniform distribution.

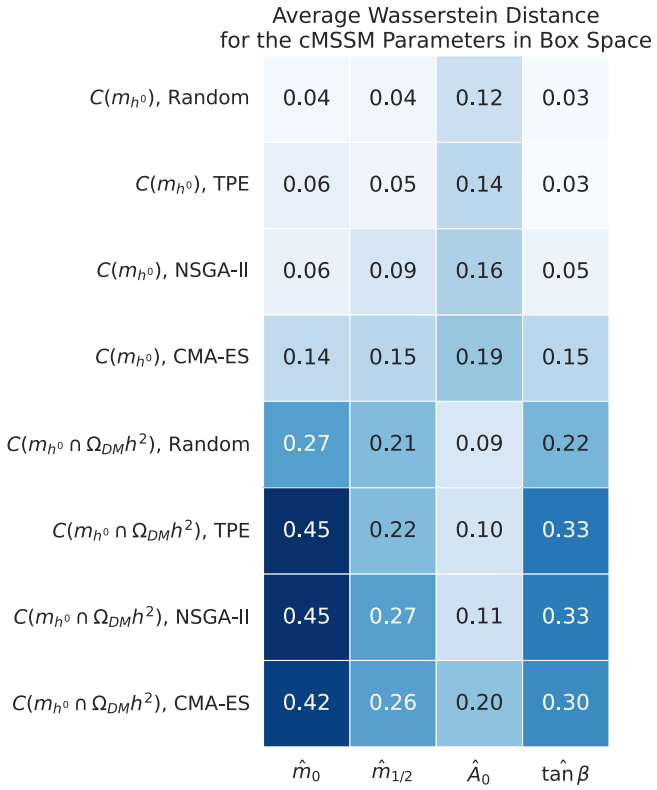


FIG. 3. Episode average of the Wasserstein distance computed on valid points for each (boxed) parameter for each sampler for the cMSSM scans.

As expected, the random sampler is the one that is closest to produce uniform distributions for the parameters, where the deviations between the resulting parameter distributions from the uniform distributions result from the constraint functions. For the other samplers, the higher values of the Wasserstein distance is a result of the sampling algorithm, given the differences in the way each sampler dynamically looks for valid points. We note that for the scan constrained only by the Higgs mass, the CMA-ES sampler considerably distorts, not only the distributions related to m_0 and A_0 in a far more pronounced manner than the remainder, indicating that it attempts to exploit the relations between these parameters and the Higgs mass, but also the distributions of the other parameters, $m_{1/2}$ and $\tan\beta$. This is in agreement with how CMA-ES works, by exploiting the statistics of the best points to sample new points close by. For the case with dark matter relic density constraint, we notice that all samplers noticeably distort the m_0 and A_0 distributions, as well as the distribution of $m_{1/2}$, a parameter that directly affects the neutralino mass spectrum and therefore dark matter relic density values. In this case, CMA-ES shows a further distortion when compared to the other samplers,

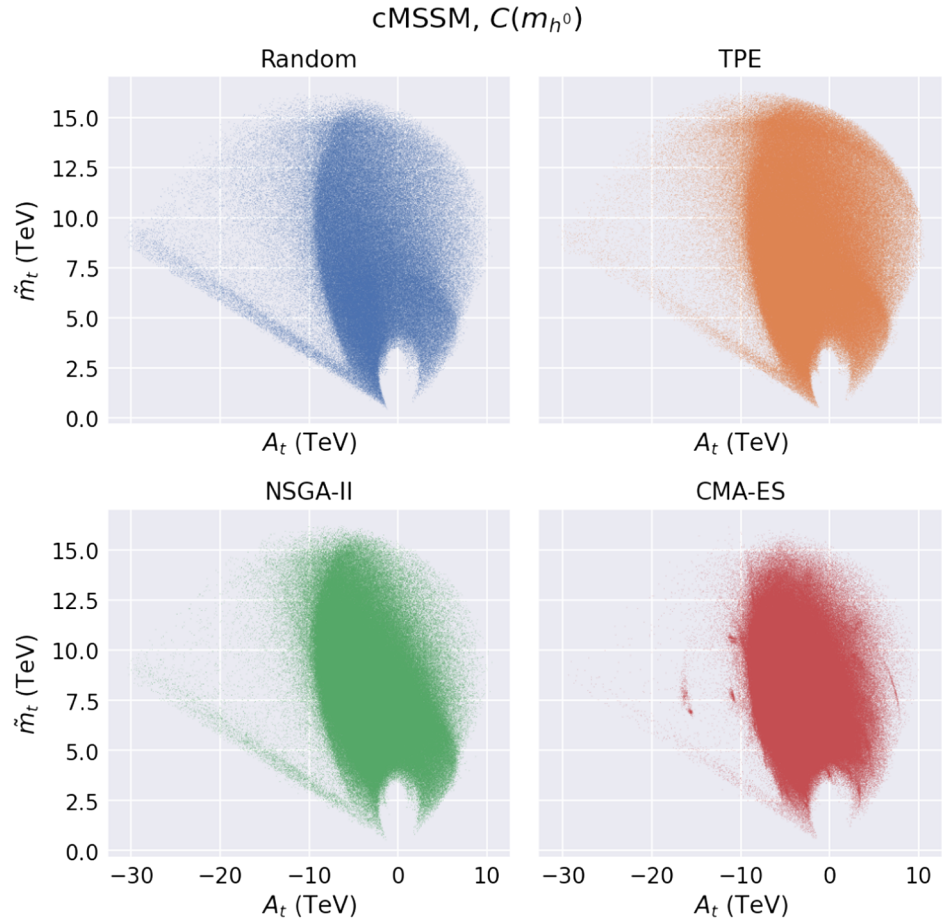


FIG. 4. $(\tilde{m}_t = \sqrt{m_{\tilde{t}_1} m_{\tilde{t}_2}}, A_t)$ scatter plots of valid points for the cMSSM scan for each sampler constrained by the Higgs mass.

mainly in $\tan\beta$, which is associated to the excess of points with lighter Higgs mass and higher dark matter relic density observed above.

Another way to look into the differences in parameter distributions across the samplers is to look into scatter plots of relevant pairs of parameters. In Fig. 4 we show the $(\tilde{m}_t = \sqrt{m_{\tilde{t}_1} m_{\tilde{t}_2}}, A_t)$ scatter plot for the cMSSM constrained only by the Higgs mass, which is parametrically dependent on these cMSSM parameters. We observe that the random sampler has the widest area coverage, specially in comparison with CMA-ES, which presents a deficit of points in the $A_t < 0$ region, while presenting various disjoint regions of high density of points resulting from the way it samples new points from a highly localized multivariate normal distribution. We also notice how the TPE covers the same region with fairly uniform density, whereas NSGA-II was capable of identifying the $\tilde{m}_t \propto -1/3A_t$ region with higher density than the other two nonrandom samplers. In this region we have rather large left-right mixing in the stop sector which enhances the corrections to the Higgs mass. Moreover, in this region there is a partial cancellation between the electroweak 1-loop contributions and the stop 2-loop contributions due to the stops and gluinos. We note for completeness that the reason of the preference of negative values for A_t is pure RGE effect as $A_t \simeq -2m_{1/2} - 0.2A_0$ for small $\tan\beta$, see, e.g., [35] and references therein.

In Fig. 5 we can observe how these scatter plots change once we include the dark matter relic density constraint. In this scan, which is far more difficult than the one without this extra constraint, we can observe new features which highlight the differences between the different samplers. First, we see that the three nonrandom samplers produced greater densities than the random sampler. Second, we can observe artefacts in the NSGA-II scatter where there is an emerging texture of vertical strips of higher density. This is a known result of genetic algorithms, where new suggested points inherit values from their parents, which can lead to the same value to be reused over many generations.[14] This happens as genetic algorithms effectively work by *swapping and combining* values of parameters between points, which leads to some combinations to be favored and survive multiple generations producing these strips. Finally, again in the

[14]In the genetic algorithms literature, recurrent combinations that survive through generations are called schema.

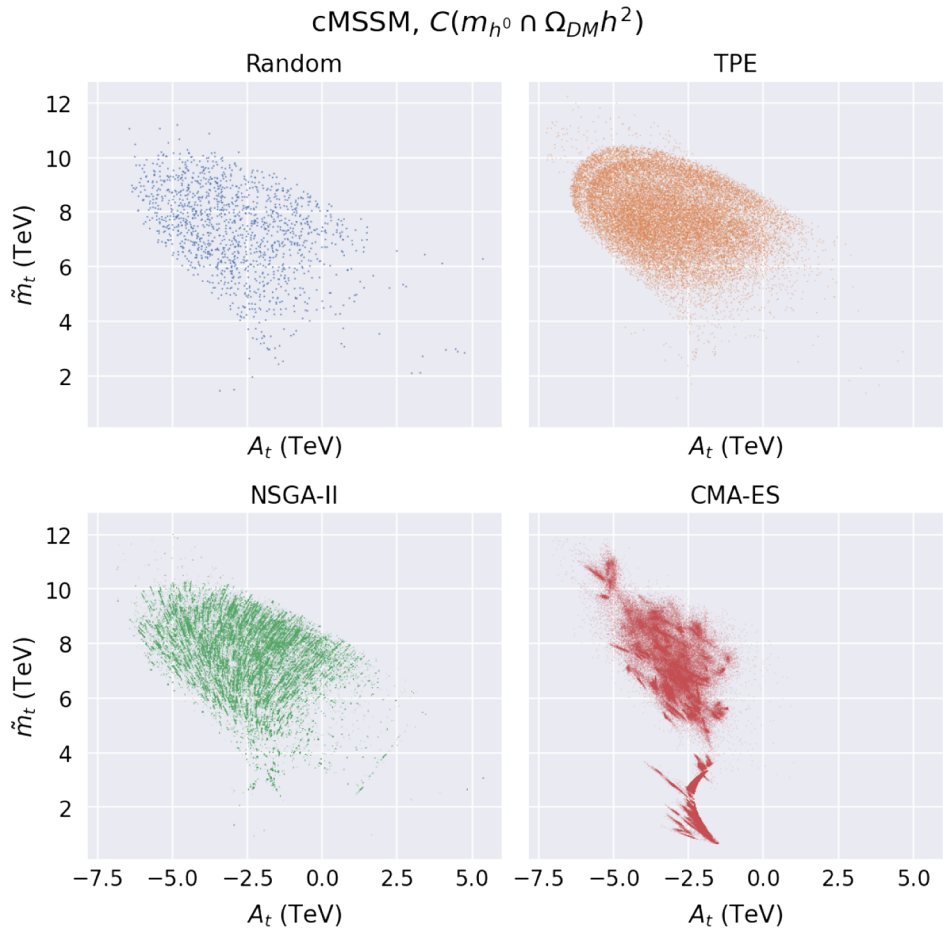


FIG. 5. $(\tilde{m}_t = \sqrt{m_{\tilde{t}_1} m_{\tilde{t}_2}}, A_t)$ scatter plots of valid points for the cMSSM scan for each sampler constrained by the Higgs mass and the dark matter relic density.

CMA-ES we observe many smaller regions of high density, which are explained by the nature of the sampler itself, since it *eagerly* samples from a multivariate normal with rolling statistics of the best points, i.e., it *exploits* the learned statistics of a local population, producing many valid points in the vicinity of the rolling mean of the best points. Due to the *eager* nature of the CMA-ES, we can also observe how it fails to capture all regions of valid points away from the *easier* region, while producing highly condensed regions of points where other samplers have only found a few, for example on the lower left quadrant, which is associated with a lighter stop, and therefore yields a lighter Higgs, as already expanded above. This region corresponds to the region where $M_1 \lesssim 1$ TeV $\mu > M_1$ in the corresponding plot of Fig. 6.

With the dark matter relic constraint it is informative to look at the (μ, M_1)[15] scatter plots as these are the relevant parameters for dark matter phenomenology. These are presented in Fig. 6. In the region with $M_1 \sim \mu \lesssim 1$ TeV one finds a mixed bino-Higgsino dark matter whereas for $\mu \sim 1$ one the dark matter is Higgsino-like. In the region $2 \leq M_1 \lesssim 3$ TeV one has a bino dark matter where the main dark matter annihilation is via a pseudoscalar Higgs funnel. The region with $M_1 \lesssim 1$ TeV $\ll \mu$ features a light stau allowing for coannihilation to obtain the correct relic density. In this region the Higgs mass is close to the lower bound which is the reason for the enhancement of the low-mass bin in case of the CMA-ES, see Fig. 2. Again, we see how the nonrandom samplers produce far denser regions of valid points, while still struggling to cover the parameter space the same way as the random sampler. However, both the TPE and the NSGA-II reproduce the overall features of the region obtained by the random sampler, whereas CMA-ES exhibits again its *eager* nature, e.g., we can see small patches of high density arising in the $M_1 \gtrsim \mu$ region. Interestingly enough, whereas all samplers discovered multiple disconnected regions, providing some evidence that these samplers can find multimodal solutions, CMA-ES has explored a particular region far more extensively than the others: the region $\mu > M_1$.

[15]We omit the equivalent scatter with M_2 as in the cMSSM $M_1 \sim M_2$ and therefore this plot provides no new insight.

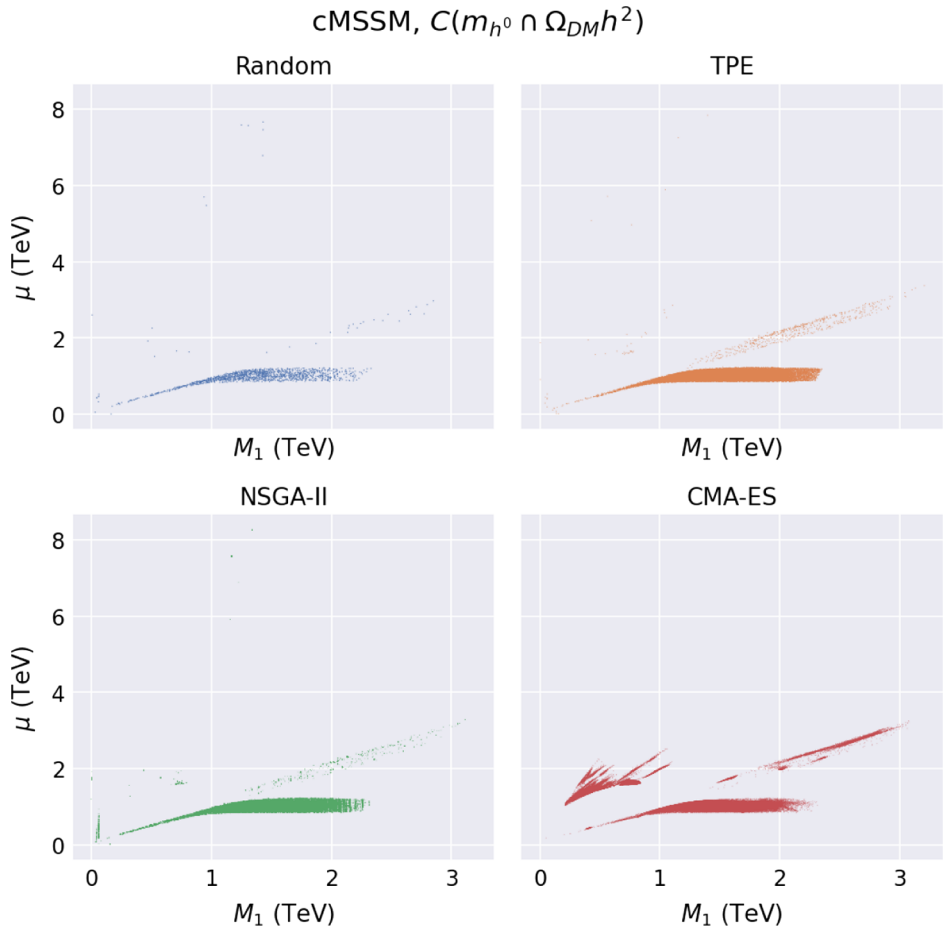


FIG. 6. (μ, M_1) scatter plots of valid points for the cMSSM scan for each sampler constrained by the Higgs mass and the dark matter relic density.

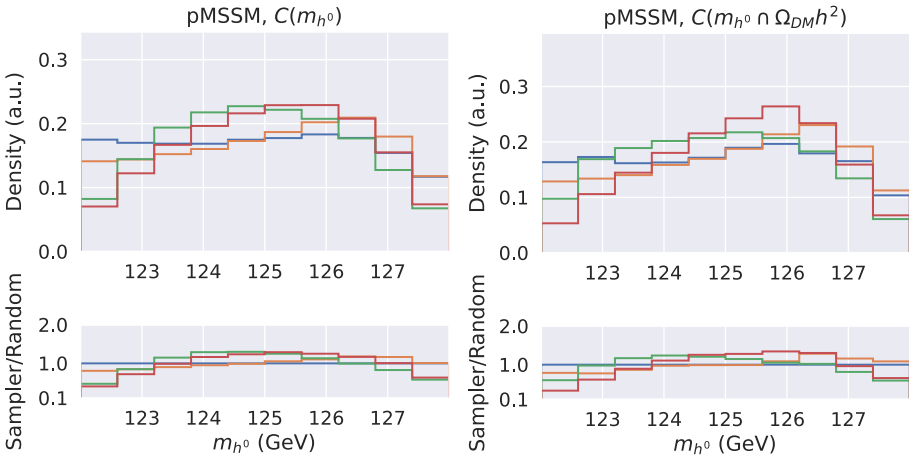
2. pMSSM

We now turn to the pMSSM. Given that our pMSSM scan covers 19 parameters, as opposed to the four parameters of the cMSSM, this scan will allow us to study the impact of increasing the dimensionality of the parameter space in the performance and results of different samplers.

In Fig. 7 we present the resulting distributions for the Higgs mass and the dark matter relic density for both pMSSM scans. Similarly to the cMSSM scans, most of nonrandom samplers focus their valid points in the interior region of the allowed interval for each observable, with the TPE being the sampler that produces distributions more similar to the random sampler. However, in contrast with the cMSSM, CMA-ES no longer seems to produce most of its valid points in close to the edges of the valid region.

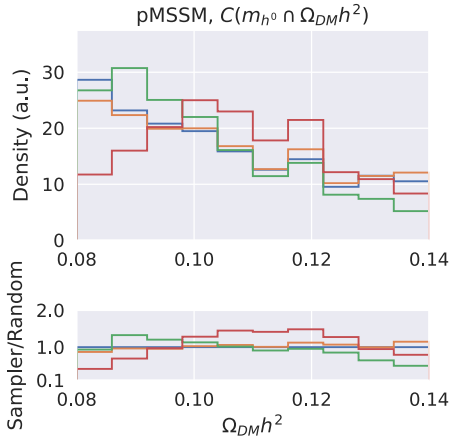
As with the cMSSM, we omit the distributions of the parameters in this section for the sake of a light discussion and instead we present the episode average Wasserstein distance for the parameters of the pMSSM scans in Fig. 8. The distributions for all pMSSM parameters can be found in the git code repository.

Just like in the cMSSM case, the random sampler produces the smallest deviations from the uniform distributions, due to its unmodified sampling. Next, we see that TPE produces almost no further distortions in the parameters, except for those directly related to the Higgs mass—A_t, $\tilde{m}_{Q_3}$, $\tilde{m}_{u_3}$—as well as, to a lesser extent, $\tilde{m}_{e_1}$, $\tilde{m}_{e_3}$, $\tilde{m}_{L_3}$, and m_{Q_1}, for the scan without the dark matter relic density constraint. When switching on the dark matter relic density constraint, the TPE produces further distortions in the parameters associated with dark matter phenomenology, namely M_1, M_2, μ. In addition the slepton mass parameters are distorted as the coannihilation channels become important if the mass difference between sleptons and neutralinos becomes sufficiently small and if the lightest neutralino has a sizeable bino-component. Similarly the enhanced distortion for the third generation squarks occurs due to the part of parameter region where there is a stop-neutralino coannihilation if the lightest neutralino has a sizeable



(a) Higgs Mass distribution for the scan with the Higgs mass constraint.

(b) Higgs Mass distribution for the scan with the Higgs mass and dark matter relic density constraints.



(c) Dark matter relic density distribution for the scan with the Higgs mass and dark matter relic density constraints.

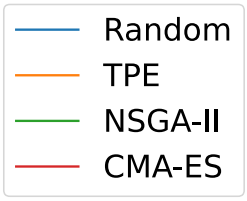


FIG. 7. Top panels: target observables distributions for the pMSSM scans. The resulting valid points histograms for each sampler are produced from joining all the episodes. Bottom panels: the ratio between the histogram of the random sampler with the remaining samplers. In all cases the histograms represent a density, which the area equals to one.

Higgsino component. Unsurprisingly, the CMA-ES is the sampler that produces the most different parameter distributions due to its *eager* nature of suggesting new points from a multivariate normal distribution around the best points. The fact that the TPE does not distort the distributions more is somehow surprising, as it makes use of Gaussian mixture models, a density learning algorithm that can be prone to the curse of dimensionality, whereas genetic algorithms such as NSGA-II are robust against this problem as they are not reliant on a learnable model.

We further investigate the impact that different samplers can have on the parameter distributions by looking at a selection of scatter plots. In Fig. 9 we present the $(A_t, \tilde{m}_t)$ scatter plot for the pMSSM scan constrained by the Higgs mass, where we can see that TPE is covering the same region as the random sampler with fairly constant point density. Furthermore, we can identify

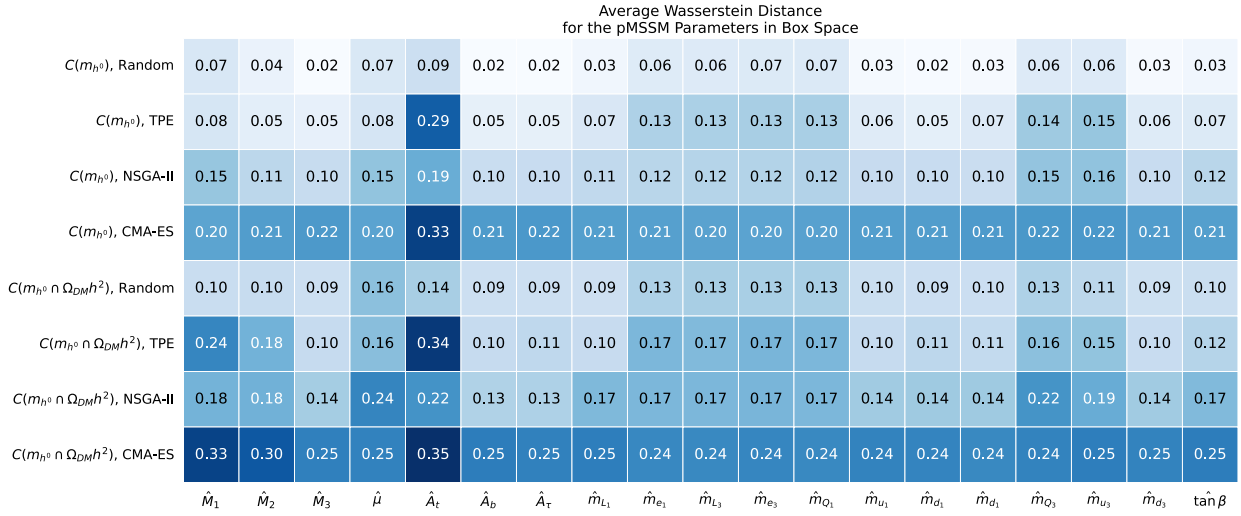


FIG. 8. Episode average of the Wasserstein distance computed on valid points for each (boxed) parameter for each sampler for the pMSSM scans.

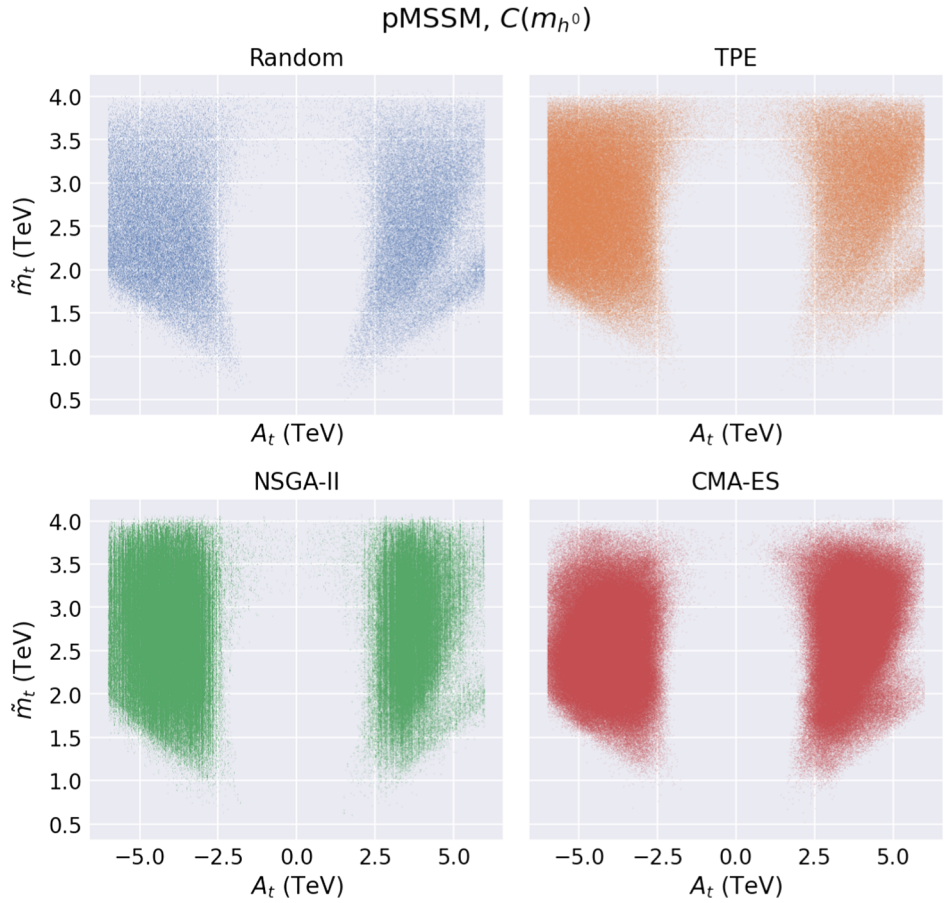


FIG. 9. $(A_t, \tilde{m}_t)$ scatter plot of valid points for the pMSSM scan for each sampling algorithm constrained by the Higgs mass.

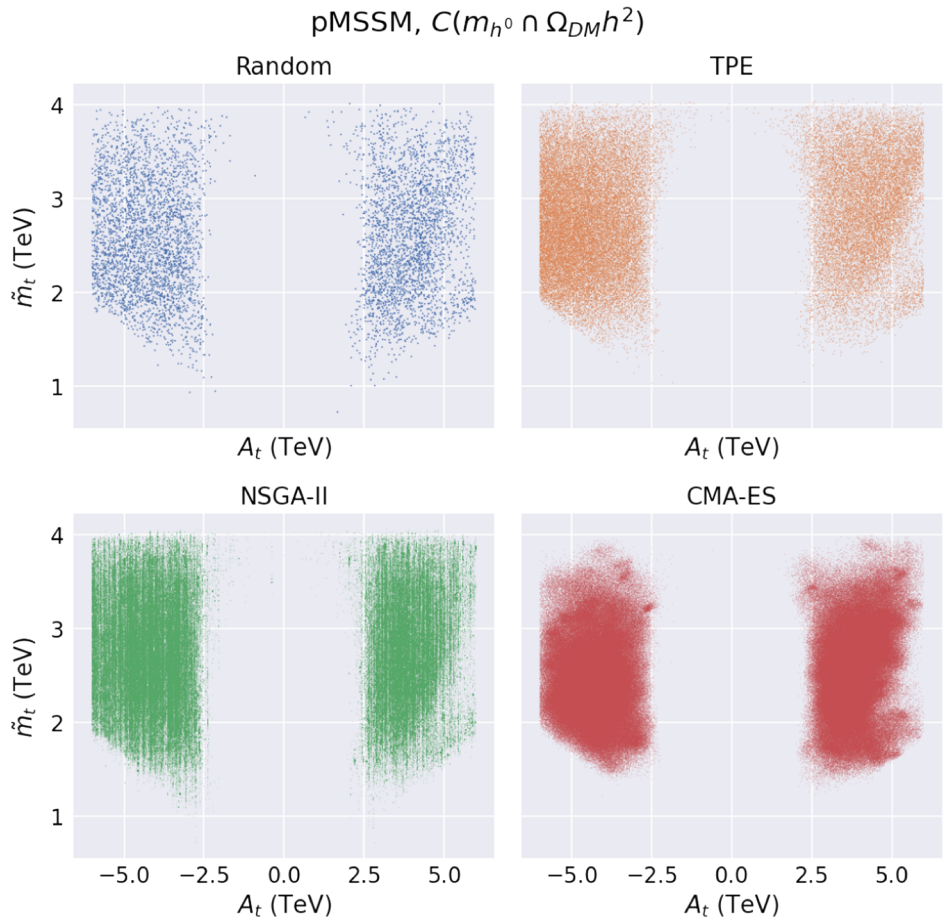


FIG. 10. $(A_t, \tilde{m}_t)$ scatter plot of valid points for the pMSSM scan for each sampling algorithm constrained by the Higgs mass and the dark matter relic density.

once again NSGA-II artefacts by noticing the emergence of strips of higher density, associated with the nature of how genetic algorithms produce new points via offspring. Finally, we see how CMA-ES focuses on easy regions of the parameter space due to its *eager* nature, more concretely we notice how it produces far less points in regions of small $\tilde{m}_t$ in comparison to the other samplers.

Looking at the equivalent scatter plots for the scan with the dark matter relic density included in Fig. 10, we observe similar features and behaviors. With special highlights to how the CMA-ES presents again smaller oval regions of higher density and the clear strips of higher density in the NSGA-II scatter, while the TPE produces a very similar result to the random sampler.

Continuing the discussion of the pMSSM with dark matter relic density constraints, we now focus on the (μ, M_1) and (μ, M_2) scatter plots in Figs. 11, 12. Some interesting features emerge in these scatter plots. We notice how the TPE is very similar to the random sampler, including the higher density regions of $\mu \gtrsim 1$ TeV. We also observe the high density strips artefacts in the NSGA-II scatters, originating from the schemas surviving multiple generations producing clustered values for the parameters. On the other hand, CMA-ES does not cover the same space as the other samplers, and, outside of the of $\mu \gtrsim 1$ TeV regions, once again produces patchy regions of higher density.

B. Efficiency and sampling metrics

Having discussed the impact of each sampler in the final parameter distributions in the previous section, in this section we compare the different samplers with respect to their efficiency and other sampling metrics.

In Fig. 13 we can see the scatter plots for the cMSSM scans, with and without the dark matter relic density constraint, for both efficiency vs episode mean Euclidean distance and efficiency vs episode total—i.e., summed over all the parameters—Wasserstein distance. These highlight the *exploration-exploitation* trade-off, as the most efficient sampler, CMA-ES, provides the worst distance metrics in accordance to the discussion from

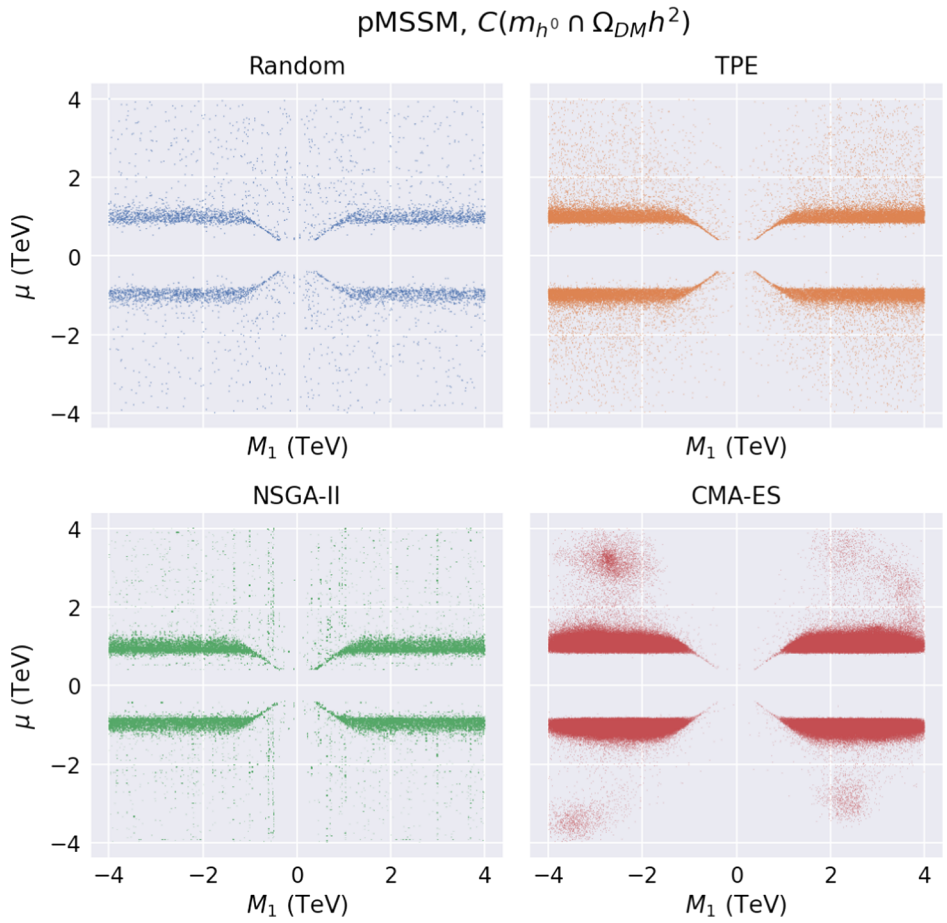


FIG. 11. (μ, M_1) scatter plot of valid points for the pMSSM scan for each sampler constrained by the Higgs mass and the dark matter relic density.

the previous section. In general, TPE provides the best parameter space coverage with slight less efficiency than the NSGA-II, which produces points which are more clustered together, as we have seen before with the strips of higher density. The NSGA-II episodes have a wider spread of possible values for the Wasserstein distance, providing a good trade-off between coverage and efficiency. We also notice that for the dark matter relic density scan, we gain at least a factor of 10 in parameter sampling efficiency, with CMA-ES increasing efficiency even further. It is also worth noticing that, in dark matter relic density constraint, the random sampler presents a significantly low efficiency of $\sim 10^{-3}$, which for our scan means that only a few ($\lesssim 10$) valid points are being sampled in each episode, causing the efficiency of each episode to be a multiple of 5×10^{-4}. That is the reason we observe the horizontal stripes for the random sampler in Fig. 13. Interestingly, we observe that for the cMSSM without dark matter relic density constraint, TPE produces on average episode mean Euclidean distances extremely close to the ones from the random sampler. This might indicate that TPE, which makes use of clustering points via a Gaussian mixture model, is sampling from far disjoint patches of the parameter space, increasing the mean Euclidean distance within the episodes. This indicates that episode mean Euclidean distance might not always be the appropriate metric for parameter space coverage.

In Fig. 14 we present the equivalent plots for the pMSSM scans, where we can observe similar trends and behaviors. Since the pMSSM enjoys greater parametric freedom than the cMSSM, the random sampler has higher sampling efficiency in the case where we consider the dark matter relic density constraint, and there is therefore slightly less room for improvement when comparing to the cMSSM case. However, it is still noticeable that the nonrandom samplers always improve parameter efficiency, with NSGA-II and CMA-ES already close to the unity efficiency.

InTables IV to VI we present the resulting statistics across the different metrics over the episodes. In Table IV we see that the random sampler has the worst efficiency across all samplers and across all physics cases. For the cases with dark matter relic density constraints, nonrandom

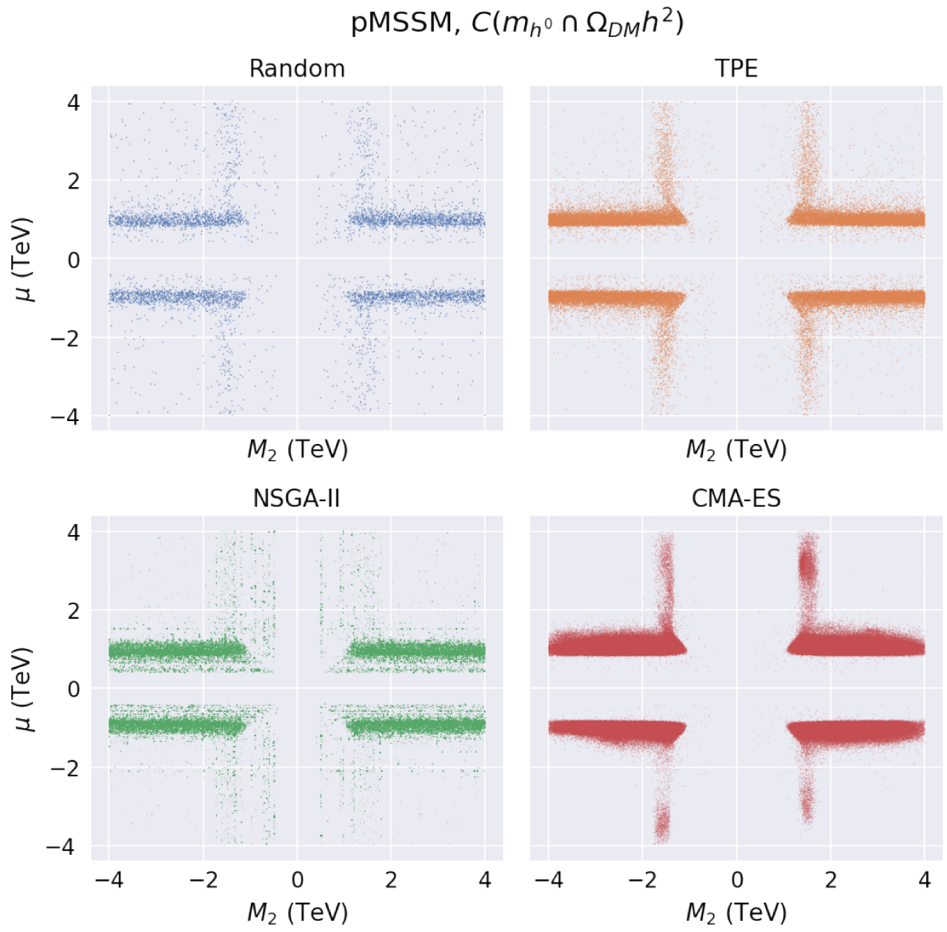
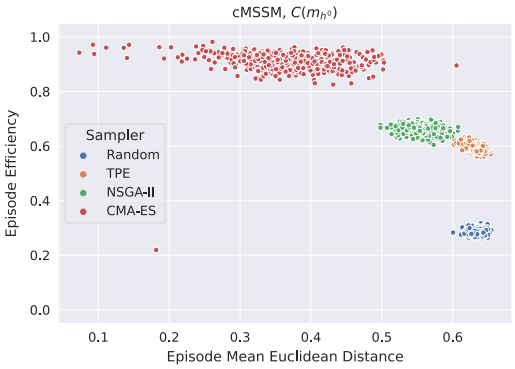


FIG. 12. (μ, M_2) scatter plot of valid points for the pMSSM scan for each sampler constrained by the Higgs mass and the dark matter relic density.

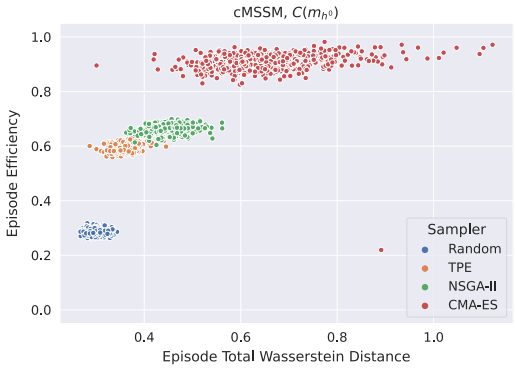
samplers can provide orders of magnitude better sampling efficiency in the best cases, and in the worst case it still more than doubles parameter space efficiency. In all cases CMA-ES is the sampler that provides the greatest efficiency, although NSGA-II comes extremely close to surpassing it in the last case. This tight contest can be explained by the large dimensionality of the pMSSM parameter space combined with the additional constrain of the dark matter relic density, where the CMA-ES sampler struggles to learn the statistics of the valid points due to the *curse of dimensionality*, which plagues shallow machine learning components. On the other hand, the NSGA-II sampler does not have any learnt component, making it scale better with the dimension of the parameter space.

Although efficiency is important, we also want to guarantee that the nonrandom samplers are properly covering the whole parameter space. In Table V we can see the average of the mean Euclidean distances. As expected, the random sampler provides the greater mean Euclidean distance, meaning that it produces valid points which are quite far apart from each other as a result of the breadth of its sampling. However, in the cMSSM without dark matter relic density constraint case, TPE comes extremely close to surpassing the random sampler in this metric. This can be due to the Gaussian mixture model sampling from two far away centers, even though the result is similar to the random case within the statistical uncertainties. In general we see that the CMA-ES produce points which are very closely together, a result due to its *eager* nature.

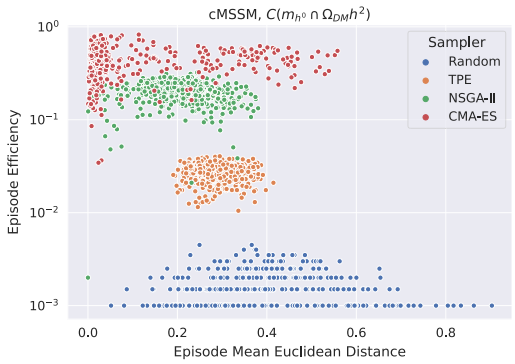
Regarding the Wasserstein distance statistics in Table VI, we observe similar trends. I.e., with the exception of the cMSSM with dark matter constraint scenario where TPE has a slightly better outcome for this metric, the random sampler is the sampler that provides the widest coverage of the parameter space as it is the one producing parameter distributions closer to a uniform distribution. This exception might be explained by the extremely low efficiency of the random sampler in this scenario, where most episodes fail to find even more than 10 valid points during the scan, which is prone to increase the variance for the Wasserstein distance metric due to low statistics. The sampler that produces the most distorted distributions is



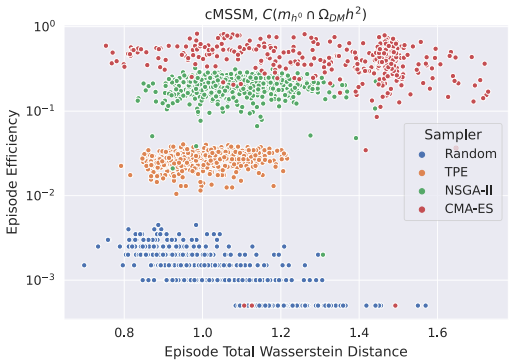
(a) Mean Euclidean distance vs efficiency per episode, for the cMSSM scan with Higgs mass constraint.



(b) Wasserstein distance vs efficiency per episode, for the cMSSM scan with Higgs mass constraint.



(c) Mean Euclidean distance vs efficiency per episode, for the cMSSM scan with Higgs mass and dark matter relic density constraints.



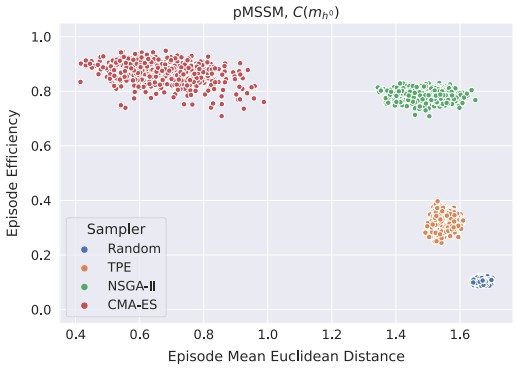
(d) Wasserstein distance vs efficiency per episode, for the cMSSM scan with Higgs mass and dark matter relic density constraints.

FIG. 13. Efficiency vs distance metrics, computed using valid points, scatter plots for each sampler for the cMSSM scans.

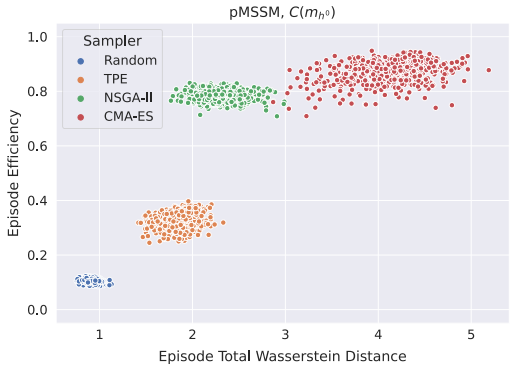
CMA-ES, a phenomenon linked to its cluster points highlighted in the table above, with TPE providing on average the best coverage out of the nonrandom samplers. NSGA-II appears just behind TPE, namely it provides similar results to TPE in the cases where the dark matter relic density is switched on, although the presence of schemas in the population prevents it from exploring the parameter space as much.

Another important aspect to compare different samplers is to see how fast they converge to valid regions, as the nonrandom samplers work sequentially, improving the quality of a suggested point with respect to the points it has suggested before. In order to assess this, we present in Fig. 15 the rolling average values for the loss, cf. Eq. (5), and the efficiency as a function of the number of trials.

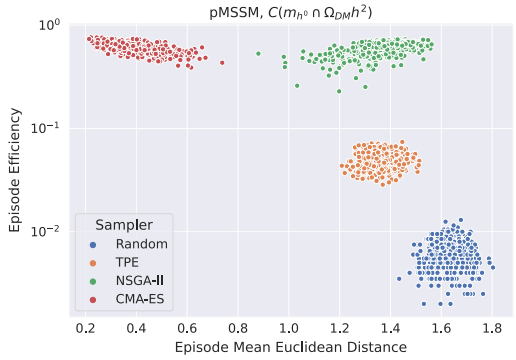
In Fig. 15(a) we see that the random sampler average loss value is constant over time. This is expected, as each sampled point of the random sampler is independent of any other sampled point. The same is not the case for the nonrandom samplers, as they attempt to produce ever better points that minimize the loss. This is explicitly observable in these plots, as we see the average loss decreasing considerably after just a few trials. Indeed, for most cases the average loss stabilizes just after a few trials, and always below the average loss of the random sampler, showing how these samplers keep producing points which are on average better than those sampled by the random sampler. The CMA-ES presents the most different behavior, with a rapid dip followed by an increase of the average loss in all cases except for the pMSSM with the dark matter relic



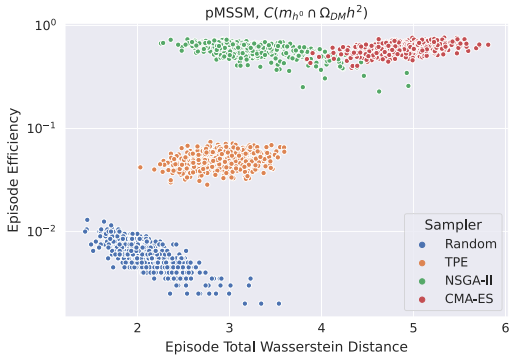
(a) Mean Euclidean distance vs efficiency per episode, for the pMSSM scan with Higgs mass constraint.



(b) Wasserstein distance vs efficiency per episode, for the pMSSM scan with Higgs mass constraint.



(c) Mean Euclidean distance vs efficiency per episode, for the pMSSM scan with Higgs mass and dark matter relic density constraints.



(d) Wasserstein distance vs efficiency per episode, for the pMSSM scan with Higgs mass and dark matter relic density constraints.

FIG. 14. Efficiency vs distance metrics, computed using valid points, scatter plots for each sampler for the pMSSM scans.

density constraint. This behavior is understood as we switch on the `restart_strategy` flag, which will restart the population once it is seemingly in a local minima in order to increase exploration. For the pMSSM with the dark matter relic density constraint case, we observe that the CMA-ES does not seem to converge within the 2000 trials allowance, and keeps suggesting better solutions. This can be due to the fact that CMA-ES works with a multivariate normal from which it samples points in this highly dimensional space, and therefore is challenged by

TABLE IV. Efficiency statistics for each sampler. The central value and the standard deviation are computed across the episodes. In bold we highlight the best nonrandom sampler for each physics case.

		Sampler			
Model	Constraint	Random	TPE	NSGA-II	CMA-ES
cMSSM	m_{h^0}	0.286 ± 0.01	0.591 ± 0.013	0.662 ± 0.015	0.909 ± 0.041
	$m_{h^0} \cap \Omega_{\mathrm{DM}}h^2$	0.001 ± 0.001	0.027 ± 0.006	0.201 ± 0.049	0.435 ± 0.157
pMSSM	m_{h^0}	0.105 ± 0.007	0.332 ± 0.03	0.786 ± 0.02	0.869 ± 0.041
	$m_{h^0} \cap \Omega_{\mathrm{DM}}h^2$	0.006 ± 0.002	0.051 ± 0.009	0.593 ± 0.083	0.605 ± 0.063

TABLE V. Mean Euclidean distance of valid points statistics for each sampler. The central value and the standard deviation are computed across the episodes. In bold we highlight the best nonrandom sampler per physics case.

		Sampler			
Model	Constraint	Random	TPE	NSGA-II	CMA-ES
cMSSM	m_{h^0}	0.634 ± 0.007	$\mathbf{0.633 \pm 0.009}$	0.554 ± 0.018	0.367 ± 0.067
	$m_{h^0} \cap \Omega_{\mathrm{DM}} h^2$	0.402 ± 0.143	$\mathbf{0.287 \pm 0.042}$	0.208 ± 0.073	0.119 ± 0.151
pMSSM	m_{h^0}	1.667 ± 0.012	$\mathbf{1.544 \pm 0.022}$	1.49 ± 0.055	0.673 ± 0.111
	$m_{h^0} \cap \Omega_{\mathrm{DM}} h^2$	1.636 ± 0.056	1.359 ± 0.053	$\mathbf{1.366 \pm 0.102}$	0.411 ± 0.089

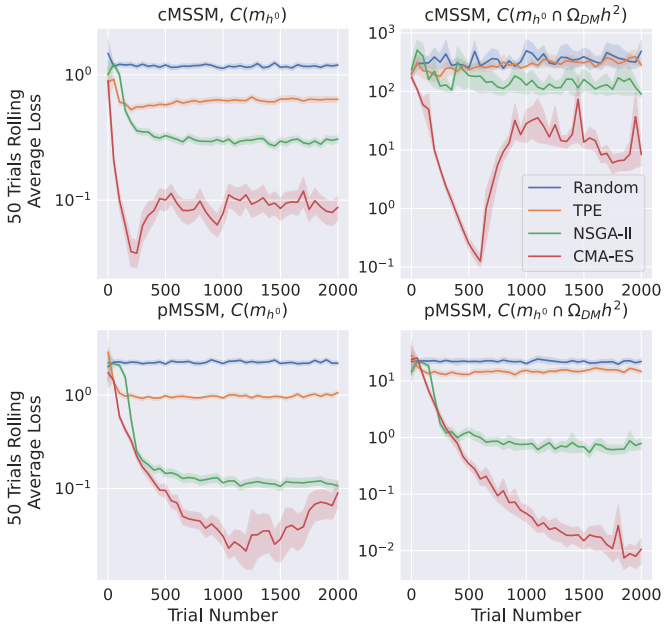
TABLE VI. Wasserstein distance computed on valid points statistics for each sampler. The central value and the standard deviation are computed across the episodes. In bold we highlight the best nonrandom sampler per physics case.

		Sampler			
Model	Constraint	Random	TPE	NSGA-II	CMA-ES
cMSSM	m_{h^0}	0.304 ± 0.013	$\mathbf{0.353 \pm 0.019}$	0.448 ± 0.033	0.662 ± 0.108
	$m_{h^0} \cap \Omega_{\mathrm{DM}} h^2$	1.034 ± 0.144	$\mathbf{1.002 \pm 0.088}$	1.088 ± 0.106	1.332 ± 0.218
pMSSM	m_{h^0}	0.907 ± 0.056	$\mathbf{1.849 \pm 0.15}$	2.322 ± 0.209	4.093 ± 0.395
	$m_{h^0} \cap \Omega_{\mathrm{DM}} h^2$	2.113 ± 0.305	$\mathbf{2.882 \pm 0.272}$	3.223 ± 0.424	4.888 ± 0.319

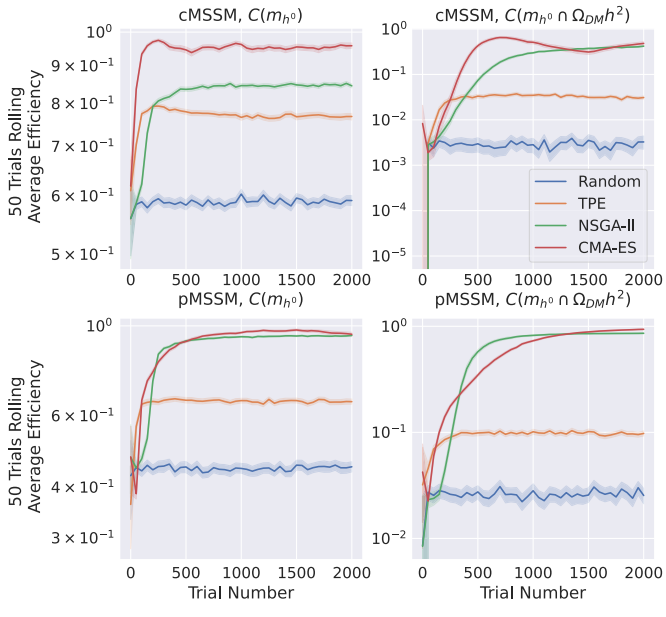
the *curse of dimensionality* as this might not be the most appropriate learnable model for such a high dimensional space.

In the plot for the rolling efficiency, Fig. 15(b), we observe a complementary behavior. All nonrandom sampler quickly saturate their sampling efficiency in almost all the cases. The exceptions are once again related to CMA-ES. For all the cases except for the pMSSM with the dark matter relic density constraint, the CMA-ES restarts its sampling after hitting an optimal sampling efficiency. For the other case, it has yet to achieve that optimal sampling efficiency point within the 2000 trials allowance within each episode.

It is interesting to point out how narrow the 95% confidence intervals are. Meaning that for sampler, each



(a) Rolling average loss per trial.



(b) Rolling average efficiency per trial.

FIG. 15. Rolling metrics history for each sampler. Each metric is computed in each episode as a function of the previous 50 trials, and the shaded bands represent 95% confidence intervals computed over the 500 episodes.

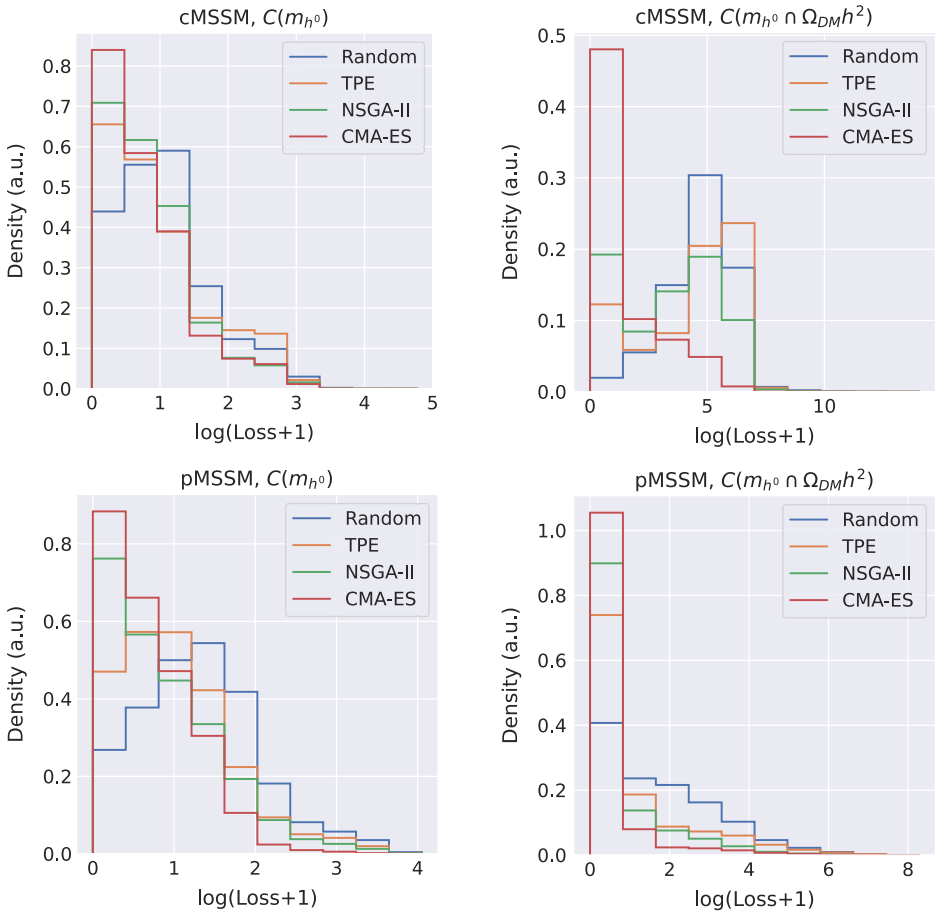


FIG. 16. Distributions of loss values for nonvalid, but physical, points for each sampler and for the different physiscs cases.

episode has a similar evolution, allowing to draw the conclusions above.

The above trial evolution plots show that the samplers progressively improve the quality of the suggested points, as measure by how likely they are to minimize the loss function. This also suggests that points that have not satisfied the conditions, but are otherwise physical (i.e., that they have successfully produced a spectrum and a dark matter candidate), should have lower loss values than points randomly sampled. In Fig. 16 we see the distribution of the values of the loss function for nonvalid, albeit physical, points. We see that for all the physics cases, the values of the losses are always lower for nonrandom samplers than for the random sampler. This is in agreement with the expectation that nonvalid points suggested by the nonrandom samplers are closer to be valid than those sampled from a random sampler.

C. Sampling time

We have already shown that the nonrandom samplers drastically improve sampling efficiency over the random sampler. However, the methodology and algorithms presented in this work are only useful if the nonrandom samplers do not impose a computational overhead that would make these scans impractically slow. In Fig. 17 we show the trial evolution time over the episodes. These plots present an artificial deformation that does not originate from our methodology: the reduction of trial time at the end of the episodes. This is due to the fact that various episodes were executed in parallel, leading to concurrency competition when reading and writing to the hard-drive, and as episodes finished it became faster to complete those still running.

In all physics cases, the random sampler is the fastest, which is expected as it does not include any new sampling algorithm. For all nonrandom sampler cases, we observe an increase of per-trial evaluation time due to the added computational overhead of the algorithm.

For all physics cases, we witness the linear growth in time for the TPE, which is in line with our expectations as the TPE fits a Gaussian mixture model that has a computational complexity that grows linearly with the number of points. This also means that the total running time of an episode, being the sum of all trials time in that episode, grows quadratically with the maximum number of trials in the episode. This is the reason why we restricted to a

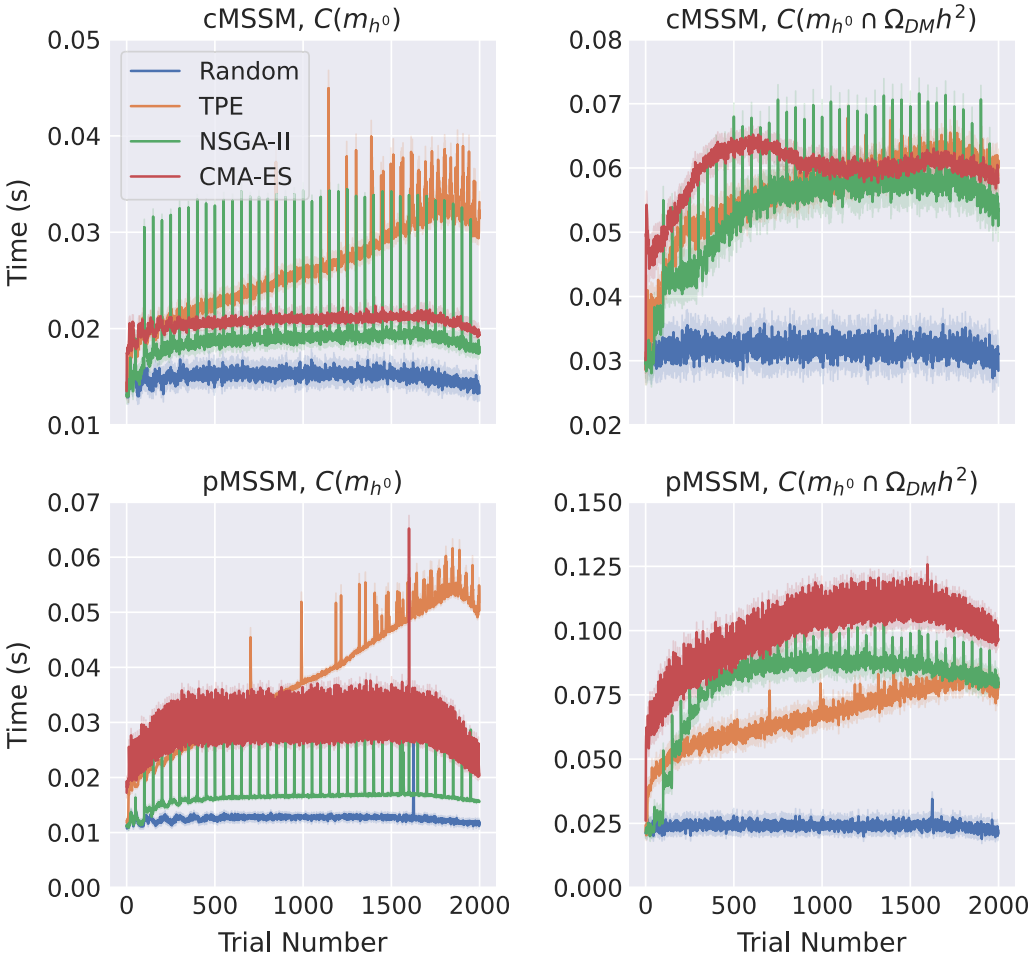


FIG. 17. Trial evolution time over the episode for each sampling algorithm. The shaded region represents 95% confidence intervals.

maximum of 2000 trials per episode, as this quadratic run-time growth, which prevents very long episodes, was identified early on in our study. This also means that for a specific problem where TPE cannot find valid points within the first few thousands trials, it will likely not be a good sampler to perform a thorough scan as its run-time will become prohibitively slow.

An interesting observation regards the spikes in time of the NSGA-II every 50 trials, giving it a comblike shape. This happens as the default population size is 50, for which after 50 trials the algorithm has to perform the genetic operations over such trials—sorting, selection, cross-over, and mutation—in order to produce candidate points to be evaluated in the following 50 trials. Despite these spikes, the NSGA-II presents the overall lightest overhead, being the fastest sampler after the random sampler in most situations.

VI. CONCLUSIONS

In this work we have reframed the parameter space scanning task for validation of BSM models as a black-box optimization problem. To accomplish this, we retain the information of an invalid point and how *far* it is from being valid using a loss function that can then be minimized using black-box optimization algorithms from the artificial intelligence and machine learning literature. We introduced three of such algorithms: tree-Parzen estimator, a Bayesian optimization algorithm; nondominated sorting genetic algorithm II, a genetic algorithm; and the covariance matrix adaptation evolution strategy, a nongenetic evolutionary algorithm. These algorithms search for valid points by interacting with the loss function, which in turn is computed using the produced observables obtained from the computational routines. In this work, we focused on the physics cases of the cMSSM and the pMSSM, with and without the further constraint of having a valid candidate for dark matter.

The novel approach presented tackles the shortcomings of current methodologies which rely on a vast collection of valid points before they can be used to sample new points, which can be a challenge for scanning tasks where random sampling can be highly inefficient from the start. Furthermore, by not being equivalent to a fit to likelihoods, our approach can be used with bounds that are derived from theory as well as experimental limits on new physics, which

are two common constraints used in BSM constraining scans that do not have a corresponding likelihood.

We showed that this approach, not requiring any *a priori* knowledge of the parameter space, provides orders of magnitude better sampling efficiencies in comparison with the random sampling strategy usually employed for this task. We showed that this benefit comes at a trade-off cost between efficiency and coverage of the parameter space, with different samplers providing distinct realizations of this trade-off: the TPE provides results similar to the random sampler, while the CMA-ES can achieve near-unity sampling efficient, and finally the NSGA-II finds its place somewhere in-between these two in terms of *exploration-exploitation* trade-off.

We have also shown how different samplers can produce different artefacts in the final distributions of the scans due to the way that they operate and sample new candidate points. Of special interest, we observed how NSGA-II and CMA-ES produced very visible artefacts in the scatter plots. In the NSGA-II case, the scatter plots presented clustered values of parameters due to the presence of schemas, which are favorable combination of parameters that can survive multiple generations. For the CMA-ES case, its sampling step based on a multivariate normal learned from the best points lead to highly dense and compact new points, which were noticeable in the scatter plots as disjoint *brush strokes*. In both cases, each algorithm produced points which were very similar to those explored thus far during the episode, leading to highly distorted distributions of the parameters, when compared to the random sampler, and larger values of the Wasserstein distance.

Furthermore, we observed how for a highly constrained scan, such as the cMSSM with dark matter relic density constraint, the CMA-ES behavior as a statistical approximation to gradient descent has produced large concentration of valid points close to the edges of the validity region. This motivates the notion of a path of least resistance in the space of the loss over the parameter space, which the other samplers, which do not operate as gradient descent, are blind to. This further suggests that different samplers traverse the parameter space differently, and how they do it will impact the resulting collection of valid points and what regions have been explored or overlooked. This motivates further dedicated work that lies outside the scope of this paper.

Ultimately, the best sampler will greatly depend on the task at hand and how difficult it is, as well as the goals of the BSM model builder in a specific study. For example, if the scan is performed on highly dimensional parameter spaces the evolutionary algorithm, NSGA-II, is better suited since it does not suffer from the *curse of dimensionality* while providing a middle ground between *exploration* and *exploitation*; if the problem revolves around a highly constrained model, where the random sampler has little efficiency, in a small dimension parameter space, then the CMA-ES would be a better choice, as it converges quickly to valid regions of the parameter space do to its *eager* nature; finally, the Bayesian algorithm, TPE, provides results more similar to the random sampler, and should therefore preferred when coverage is the main concern, although it will struggle to find good points if it fails to converge to a valid region within the first few thousand points due to its run-time becoming prohibitively slow.

Although we have shown the great potential benefit of using nonrandom samplers to perform parameter space scanning of BSM models, our work also points at future directions to improve upon the proposed methodologies. First, despite choosing some options that differ from the default parameters, we have not undertaken any optimization of the samplers, which could further improve the presented metrics. Second, we have to reiterate that the proposed algorithms were not designed for the specific case of BSM parameter space scan and constraining—which requires extensive coverage over highly multidimensional spaces—and therefore there is the potential to further improve them, or design new ones, that can mitigate the *exploration-exploitation* cost of choosing one side over the other, or the sensitivity to the *curse of dimensionality* of same of the samplers. Finally, we made an explicit choice of summing together two constraint functions instead of optimising each separately as a *multiobjective* optimization problem. This choice was made so that we could use different optimizers that cannot perform such task, such as the CMA-ES, but it is likely that algorithms like NSGA-II, which were designed especially for such problems, will provide even better samplers for problems that involve multiple joint constraints.

Finally, we notice that the methodology herein is not restricted to SUSY model building, and can be used with any computational routine and set of constraints—regardless of the BSM framework and computational language where the routines are written—and therefore provides a general new paradigm for parameter space scanning and BSM model validation.

The code of this work is available in [58].

ACKNOWLEDGMENTS

We thank José Santiago Pérez and Jorge Romão for the careful reading of the paper draft and for the useful discussions. This work is supported by FCT—Fundação para a Ciência e a Tecnologia, I. P. under project No. CERN/FIS-PAR/0024/2019. F. A. S. is also supported by FCT under the research grant with reference No. UI/BD/153105/2022. The computational work was partially done using the resources made available by Rede Nacional de Computação Avançada (RNCA) and Infraestrutura Nacional de Computação Distribuída (INCD) under project CPCA/A1/401197/2021.

[1] Matthew Feickert and Benjamin Nachman, A living review of machine learning for particle physics, arXiv:2102.02770.

[2] Jie Ren, Lei Wu, Jin Min Yang, and Jun Zhao, Exploring supersymmetry with machine learning, Nucl. Phys. **B943**, 114613 (2019).

[3] Florian Staub, xbit: An easy to use scanning tool with machine learning abilities, arXiv:1906.03277.

[4] B. S. Kronheim, M. P. Kuchera, H. B. Prosper, and A. Karbo, Bayesian neural networks for fast SUSY predictions, Phys. Lett. B **813**, 136041 (2021).

[5] Sascha Caron, Tom Heskes, Sydney Otten, and Bob Stienen, Constraining the parameters of high-dimensional models with active learning, Eur. Phys. J. C **79**, 944 (2019).

[6] Mark D. Goodsell and Ari Joury, Active learning BSM parameter spaces, arXiv:2204.13950.

[7] Jacob Hollingsworth, Michael Ratz, Philip Tanedo, and Daniel Whiteson, Efficient sampling of constrained high-dimensional theoretical spaces with machine learning, Eur. Phys. J. C **81**, 1138 (2021).

[8] Sascha Caron, Jong Soo Kim, Krzysztof Rolbiecki, Roberto Ruiz de Austri, and Bob Stienen, The BSM-AI project: SUSY-AI–generalizing lhc limits on supersymmetry with machine learning, Eur. Phys. J. C **77**, 257 (2017).

[9] Atılım Günes Baydin *et al.*, Differentiable programming in high-energy physics, Submitted as a Snowmass LOI, 2020, https://www.snowmass21.org/docs/files/summaries/CompF/SNOWMASS21-CompF5_CompF3_Gordon_Watts-046.pdf.

[10] Lukas Heinrich and Michael Kagan, Differentiable matrix elements with madjax, arXiv:2203.00057.

[11] Gurtej Kanwar, Michael S. Albergo, Denis Boyda, Kyle Cranmer, Daniel C. Hackett, Sébastien Racanière, Danilo Jimenez Rezende, and Phiala E. Shanahan, Equivariant Flow-Based Sampling for Lattice Gauge Theory, Phys. Rev. Lett. **125**, 121601 (2020).

[12] P. A. Zyla *et al.*, Review of particle physics, Prog. Theor. Exp. Phys. **2020**, 083C01 (2020).

[13] P. Slavich *et al.*, Higgs-mass predictions in the MSSM and beyond, Eur. Phys. J. C **81**, 450 (2021).

[14] Fawzi Boudjema, Guillaume Drieu La Rochelle, and Suchita Kulkarni, One-loop corrections, uncertainties and approximations in neutralino annihilations: Examples, Phys. Rev. D **84**, 116001 (2011).

[15] F. Boudjema, G. Drieu La Rochelle, and A. Mariano, Relic density calculations beyond tree-level, exact calculations versus effective couplings: The ZZ final state, Phys. Rev. D **89**, 115020 (2014).

[16] J. Harz, B. Herrmann, M. Klasen, K. Kovarik, and P. Steppeler, Theoretical uncertainty of the supersymmetric dark matter relic density from scheme and scale variations, Phys. Rev. D **93**, 114023 (2016).

[17] Jordan Bernigaud, Adam K. Forster, Björn Herrmann, Stephen F. King, Werner Porod, and Samuel J. Rowley, Data-driven analysis of a SUSY GUT of flavour, J. High Energy Phys. 05 (2022) 156.

[18] Kyle Cranmer *et al.*, Publishing statistical models: Getting the most out of particle physics experiments, SciPost Phys. **12**, 037 (2022).

[19] Logan Morrison, Stefano Profumo, and John Tamanas, Simulation based inference for efficient theory space sampling: An application to supersymmetric explanations of the anomalous muon $g-2$, Phys. Rev. D **106**, 115016 (2022).

[20] Georges Aad *et al.*, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B **716**, 1 (2012).

[21] Serguei Chatrchyan *et al.*, Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, Phys. Lett. B **716**, 30 (2012).

[22] Gordon L. Kane, Christopher F. Kolda, Leszek Roszkowski, and James D. Wells, Study of constrained minimal supersymmetry, Phys. Rev. D **49**, 6173 (1994).

[23] A. Djouadi *et al.*, The minimal supersymmetric standard model: Group summary report, arXiv:hep-ph/9901246.

[24] J. E. Camargo-Molina, B. O'Leary, W. Porod, and F. Staub, Stability of the CMSSM against sfermion VEVs, J. High Energy Phys. 12 (2013) 103.

[25] J. E. Camargo-Molina, B. Garbrecht, B. O'Leary, W. Porod, and F. Staub, Constraining the Natural MSSM through tunneling to color-breaking vacua at zero and non-zero temperature, Phys. Lett. B **737**, 156 (2014).

[26] Leszek Roszkowski, Enrico Maria Sessolo, and Sebastian Trojanowski, WIMP dark matter candidates and searches—current status and future prospects, Rep. Prog. Phys. **81**, 066201 (2018).

[27] Kim Griest and David Seckel, Three exceptions in the calculation of relic abundances, Phys. Rev. D **43**, 3191 (1991).

[28] Werner Porod, SPheno, a program for calculating supersymmetric spectra, SUSY particle decays and SUSY particle production at $e^+ e^-$ colliders, Comput. Phys. Commun. **153**, 275 (2003).

[29] W. Porod and F. Staub, SPheno 3.1: Extensions including flavour, *CP*-phases and models beyond the MSSM, Comput. Phys. Commun. **183**, 2458 (2012).

[30] G. Belanger, F. Boudjema, A. Pukhov, and A. Semenov, Dark matter direct detection rate in a generic model with micrOMEGAs 2.2, Comput. Phys. Commun. **180**, 747 (2009).

[31] G. Belanger, F. Boudjema, and A. Pukhov, micrOMEGAs: A code for the calculation of dark matter properties in generic models of particle interaction, arXiv:1402.0787.

[32] Peter Z. Skands *et al.*, SUSY Les Houches accord: Interfacing SUSY spectrum calculators, decay packages, and event generators, J. High Energy Phys. 07 (2004) 036.

[33] B. C. Allanach *et al.*, SUSY les Houches accord 2, Comput. Phys. Commun. **180**, 8 (2009).

[34] Florian Staub and Werner Porod, Improved predictions for intermediate and heavy supersymmetry in the MSSM and beyond, Eur. Phys. J. C **77**, 338 (2017).

[35] G. A. Blair, W. Porod, and P. M. Zerwas, The reconstruction of supersymmetric theories at high-energy scales, Eur. Phys. J. C **27**, 263 (2003).

[36] B. C. Allanach, A. Djouadi, J. L. Kneur, W. Porod, and P. Slavich, Precise determination of the neutral Higgs boson masses in the MSSM, J. High Energy Phys. 09 (2004) 044.

[37] Gerard Jungman, Marc Kamionkowski, and Kim Griest, Supersymmetric dark matter, Phys. Rep. **267**, 195 (1996).

[38] John Ellis and Keith A. Olive, Revisiting the Higgs mass and dark matter in the CMSSM, Eur. Phys. J. C **72**, 2005 (2012).

[39] John R. Ellis, Toby Falk, Keith A. Olive, and Mark Srednicki, Calculations of neutralino–stau coannihilation channels and the cosmologically relevant region of MSSM parameter space, Astropart. Phys. **13**, 181 (2000); **15**, 413(E) (2001).
[40] Celine Boehm, Abdelhak Djouadi, and Manuel Drees, Light scalar top quarks and supersymmetric dark matter, Phys. Rev. D **62**, 035012 (2000).
[41] Thomas J. LeCompte and Stephen P. Martin, Compressed supersymmetry after 1 fb^{-1} at the Large Hadron Collider, Phys. Rev. D **85**, 035023 (2012).
[42] CMS Collaboration, Search for supersymmetry in proton-proton collisions at 13 TeV in final states with jets and missing transverse momentum, J. High Energy Phys. 10 (2019) 244.
[43] ATLAS Collaboration, Search for squarks and gluinos in final states with jets and missing transverse momentum using 139 fb^{-1} of $\sqrt{s} = 13$ TeV pp collision data with the ATLAS detector, J. High Energy Phys. 02 (2021) 143.
[44] Peter Athron, Csaba Balázs, Douglas H. J. Jacob, Wojciech Kotlarski, Dominik Stöckinger, and Hyejung Stöckinger-Kim, New physics explanations of a_μ in light of the FNAL muon $g-2$ measurement, J. High Energy Phys. 09 (2021) 080.
[45] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl, Algorithms for hyper-parameter optimization, in *Advances in Neural Information Processing Systems*, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Curran Associates, Inc., 2011), Vol. 24, https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.
[46] James Bergstra, Daniel Yamins, and David Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in *Proceedings of the 30th International Conference on Machine Learning*, edited by Sanjoy Dasgupta and David McAllester, Volume 28 of Proceedings of Machine Learning Research (PMLR, Atlanta, Georgia, USA, 2013), pp. 115–123.
[47] Yoshihiko Ozaki, Yuki Tanigaki, Shuhei Watanabe, and Masaki Onishi, Multiobjective tree-structured parzen estimator for computationally expensive optimization problems, in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference, GECCO '20* (Association for Computing Machinery, New York, NY, USA, 2020), pp. 533–541.
[48] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. **6**, 182 (2002).
[49] Nikolaus Hansen, The CMA evolution strategy: A tutorial, arXiv:1604.00772.
[50] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19* (Association for Computing Machinery, New York, NY, USA, 2019), pp. 2623–2631.
[51] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux, The NumPy array: A structure for efficient numerical computation, Comput. Sci. Eng. **13**, 22 (2011).
[52] Skipper Seabold and Josef Perktold, Statsmodels: Econometric and statistical modeling with Python, in *Proceedings of the 9th Python in Science Conference* (Austin, TX, 2010), Vol. 57, p. 61, https://pdfs.semanticscholar.org/3a27/6417e5350e29cb6bf04ea5a4785601d5a215.pdf.
[53] Pauli Virtanen *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).
[54] Jeff Reback *et al.*, pandas-dev/pandas: Pandas 1.0.3, 10.5281/zenodo.3715232 (2020).
[55] John D Hunter, Matplotlib: A 2d graphics environment, Comput. Sci. Eng. **9**, 90 (2007).
[56] Michael L Waskom, Seaborn: Statistical data visualization, J. Open Source Softwaare **6**, 3021 (2021).
[57] Andrzej Novak *et al.*, scikit-hep/mplhep: v0.3.23, 10.5281/zenodo.6332486 (2022).
[58] https://gitlab.com/lip_ml/blackboxbsm