City Research Online

City St George's, University of London

Citation: Deen, R., He, Y-H., Lee, S-J. & Lukas, A. (2022). Machine learning string standard models. Physical Review D (PRD), 105(4), 046001. doi: 10.1103/physrevd.105.046001

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/27799/

Link to published version: https://doi.org/10.1103/physrevd.105.046001

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Machine learning string standard models

Rehan Deen,[1,*] Yang-Hui He,[2,3,4,†] Seung-Joo Lee,[5,6,‡] and Andre Lukas[1,§]

[1]*Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Parks Road, Oxford OX1 3PU, United Kingdom*

[2]*School of Physics, NanKai University, Tianjin 300071, People's Republic of China*

[3]*Department of Mathematics, City, University of London, London EC1V0HB, United Kingdom*

[4]*Merton College, University of Oxford, Oxford OX1 4JD, United Kingdom*

[5]*Center for Theoretical Physics of the Universe, Institute for Basic Science, Daejeon 34051, South Korea*

[6]*CERN, Theory Department, 1 Esplande des Particules, Geneva 23, CH-1211, Switzerland*

(Received 23 September 2021; accepted 18 January 2022; published 2 February 2022)

We study machine learning of phenomenologically relevant properties of string compactifications, which arise in the context of heterotic line bundle models. Both supervised and unsupervised learning are considered. We find that, for a fixed compactification manifold, relatively small neural networks are capable of distinguishing consistent line bundle models with the correct gauge group and the correct chiral asymmetry from random models without these properties. The same distinction can also be achieved in the context of unsupervised learning, using an autoencoder. Learning nontopological properties, specifically the number of Higgs multiplets, turns out to be more difficult, but is possible using sizeable networks and feature-enhanced datasets.

DOI: 10.1103/PhysRevD.105.046001

I. INTRODUCTION

Techniques from machine learning have recently been introduced into string theory [1–5] and have been explored for a variety of different machine learning architectures and string theory settings (for reviews see Refs. [6,7] and references therein). One of the main motivations for bringing modern data science methods to string theory is the vast amount of data generated by string theory. String theory data describes different solutions to the theory and, at the most basic level of topological or quasitopological properties, is discrete data described by sets of integers, which specifies the compactification manifold and branes/fluxes thereon.

For a given string solution this integer data determines (to a large extent) the massless spectrum of the associated low-energy theory. It is widely believed that the set of (topological types of) string solutions is finite, although no general proof is known. (For instance, it is a conjecture of Yau that the topological types of compact smooth connected Calabi-Yau manifolds is finite in every dimension.) If this is indeed the case it is clear, however, that the set is vast and an impressive lower bound of $\sim 10^{272000}$ has been given in Ref. [8]. Can machine learning help to deal with this vast set of string vacua, for example by organizing the data or distinguishing between models with different properties?

One of the central problems of string theory remains understanding its relation to established low-energy particle physics. It is now known that string theory contains many models with a promising spectrum of particles and, in the context of a specific construction, a lower bound of $\sim 10^{23}$ ("a mole of models" [9]) for the number of such models has been given in Ref. [10]. Nevertheless, string solutions which lead to such models constitute a very small fraction of the total. Finding phenomenologically promising models within string theory and even deciding whether given string vacua are promising can be nontrivial tasks. In this paper we will deal with the second problem and the main question we will address is whether machine learning can distinguish string vacua which lead to phenomenologically attractive models from those which do not. In other words, can a neural network learn whether or not a string solution is of physical interest?

We will be addressing this question in the context of heterotic line bundle models [11–13], a class of models which has the virtue of being conceptually relatively simple and for which sizeable sets of phenomenologically promising models are known. This means that training sets for machine learning can readily be constructed. To be clear

*rehan.deen@physics.ox.ac.uk
†hey@maths.ox.ac.uk
‡seung.joo.lee@cern.ch
§andre.lukas@physics.ox.ac.uk

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

about terminology, by "standardlike" model (SLM for short) we mean a consistent string solution which gives rise to the standard model gauge group and has the correct chiral asymmetry of three families of quarks and leptons. Whether a string solution is a SLM is a topological question, that is, it only depends on topological quantities such as Chern classes and indices. Further requirements on a SLM toward realistic physics are the absence of vector-like matter and the presence of Higgs doublets. These properties are controlled by bundle cohomology and are less robust. They can depend on the (complex structure) moduli of the model and, in this sense, are nontopological.

For the purpose of this paper, we will only consider supervised learning and unsupervised learning with autoencoders, either based on fully connected feed-forward networks. Unsupervised learning for heterotic orbifold models with autoencoders has been studied in Ref. [14] and reinforcement learning of string models has been investigated in Refs. [15,16].

In the next section, we briefly review heterotic line bundle models and in Sec. III we describe the associated datasets. In Sec. IV we show that simple fully connected networks with supervised learning can distinguish SLMs and non-SLMs. This can also be done via unsupervised learning, using an autoencoder, as explained in Sec. V. Finally, in Sec. VI, we investigate whether neural networks can detect the presence or absence of Higgs multiplets in SLMs. We find this is possible, but requires a more careful approach which draws on previous experience with line bundle cohomology [17–20] and machine learning of line bundle cohomology formulas [21]. We conclude in Sec. VII.

II. HETEROTIC LINE BUNDLE MODELS

Heterotic line bundle models have been introduced and analyzed in Refs. [11–13] and we refer to these papers for details. Here, we present a concise summary with emphasis on aspects relevant to machine learning applications. The data which specifies a heterotic line bundle model is a tuple (X, Γ, V), where X is a Calabi-Yau three-fold, Γ is a freely acting symmetry on X and V is a (Γ-equivariant) line bundle sum on X, here taken to be of rank five[1]

$$V = \bigoplus_{a=1}^{5} L_a, \tag{2.1}$$

where $L_a \to X$ are line bundles. For this to define a consistent string compactification on X we require that

$$c_1(V) = \sum_{a=1}^{5} c_1(L_a) \stackrel{!}{=} 0,$$
$$c_2(TX) - c_2(V) \in \text{Mori cone of } X, \tag{2.2}$$

where the first condition is required for an embedding of the structure group of V into E_8 and the second condition guarantees that the anomaly cancelation condition can be satisfied. A further consistency condition, which serves to ensure that the bundle V preserves supersymmetry, is that the slopes

$$\mu(L_a) := \int_X c_1(L_a) \wedge J^2 \stackrel{!}{=} 0 \tag{2.3}$$

for all five line bundles vanish simultaneously for some Kähler class J of X.

Such a consistent model on X defines a four-dimensional $N = 1$ grand unified theory (GUT) with gauge group $SU(5)$ and matter fields in the $SU(5)$ multiplets $\mathbf{10}$, $\overline{\mathbf{10}}$, $\bar{\mathbf{5}}$ and $\mathbf{5}$ (plus fields uncharged under $SU(5)$).[2] We recall that one standard model family fits precisely into the $SU(5)$ multiplet $\mathbf{10} \oplus \bar{\mathbf{5}}$. The numbers of these multiplets are governed by the bundle cohomologies

$$\#\mathbf{10} = h^1(X, V), \qquad \#\overline{\mathbf{10}} = h^2(X, V),$$
$$\#\bar{\mathbf{5}} = h^1(X, \wedge^2 V), \qquad \#\mathbf{5} = h^2(X, \wedge^2 V). \tag{2.4}$$

The chiral asymmetry is measured by the index of the bundle V and is given by[3]

$$\#\text{GUT families} = \#\mathbf{10} - \#\overline{\mathbf{10}} = \#\bar{\mathbf{5}} - \#\mathbf{5}$$
$$= -\text{ind}(V) \stackrel{!}{=} 3|\Gamma|, \tag{2.5}$$

where $|\Gamma|$ is the order of the freely acting symmetry Γ. Note this equation implies that the chiral spectrum always consist of a number of complete GUT families $\mathbf{10} \oplus \bar{\mathbf{5}}$. The reason for demanding $3|\Gamma|$ rather than just 3 such families is that the final model is defined on the quotient Calabi-Yau $\hat{X} = X/\Gamma$ with a bundle $\hat{V} \to \hat{X}$ which descends from V. Since the index satisfies $\text{ind}(\hat{V}) = \text{ind}(V)/|\Gamma|$, Eq. (2.5) does indeed guarantee three chiral families "downstairs." The downstairs model on $\hat{X}$ also allows for the inclusion of an additional flat bundle (a Wilson line) which breaks $SU(5)$ to the standard model gauge group without disturbing the index. In the present paper, we will not be concerned with the downstairs construction. We know that it can be carried out and it is,

[1] Line bundle sums with different rank, specifically with rank four, can also lead to phenomenologically promising models. However, Wilson line breaking of such models requires large symmetries Γ which are rare. As a result, large datasets are not yet available.

[2] Additional Green-Schwarz anomalous $U(1)$ gauge symmetries are also present but their associated vector bosons are usually supermassive.

[3] This follows from $h^0(X, L) = h^3(X, L) = 0$ which holds for line bundles L with vanishing slope and from $\text{ind}(V) = \text{ind}(\wedge^2 V)$, a general property of rank five bundles V.

hence, sufficient to focus on the data of the upstairs GUT model.[4]

A further physical condition, in addition to Eq. (2.5), is

$$\#\overline{\mathbf{10}} = h^2(X,V) \stackrel{!}{=} 0, \tag{2.6}$$

which ensures the absence of vectorlike $\mathbf{10}$–$\overline{\mathbf{10}}$ pairs. We also need at least one vectorlike $\bar{\mathbf{5}}$–$\mathbf{5}$ pair to account for the Higgs, which amounts to the condition

$$\#\text{pairs of Higgs doublets} = \#\mathbf{5} = h^2(X,V) \stackrel{!}{>} 0. \tag{2.7}$$

In summary, the data (X,Γ,V) defines a consistent string compactification, iff the Chern class conditions (2.2) and the slope conditions (2.3) are satisfied. The most basic physical requirement on such a model is that is has the correct chiral asymmetry of matter which amounts to the condition (2.5). We call a model with these properties, all of which are topological, a standardlike model (SLM). The next most basic physical requirements are the absence of vectorlike pairs of matter, Eq. (2.6), and the presence of multiplets to account for the Higgs, Eq. (2.7). Both of these conditions depend on cohomology and are, hence, not strictly topological.[5] For the purpose of machine learning, we need to translate the above geometrical data into numerical, integer data. To this end, we introduce a basis $\{J_i\}$ of the second cohomology and a dual basis $\{\nu^i\}$ of the fourth cohomology of X, where $i = 1,\ldots,h$ and $h = h^{1,1}(X)$. Relative to this basis the second Chern class of the tangent bundle can be written as $c_2(TX) = c_{2i}(TX)\nu^i$, with integers $c_{2i}(TX)$. The Kähler form of X can be expanded as $J = t^i J_i$, where t^i are the Kähler parameters.[6] We will also need the triple intersection numbers of X, defined by

$$d_{ijl} = \int_X J_i \wedge J_j \wedge J_l. \tag{2.8}$$

Further, a line bundle on X can be represented by an h-dimensional integer vector $k = (k^1,\ldots,k^h)$ and is written as $\mathcal{O}_X(k)$, such that $c_1(\mathcal{O}_X(k)) = k^i J_i$. This means that the line bundle sum (2.1) takes the form

$$V = \bigoplus_{a=1}^{5} \mathcal{O}_X(k_a) \leftrightarrow K = (k_a^i) \tag{2.9}$$

and it can be represented by the $h \times 5$ integer matrix K. In practice, a model, for a fixed Calabi-Yau manifold X and a fixed symmetry Γ, will be represented by this integer matrix K. In terms of this matrix, the Chern classes of such a line bundle sum are given by[7]

$$c_1(V) = \sum_{a=1}^{5} k_a^i J_i,$$
$$c_2(V) = -\frac{1}{2} d_{ijl} \sum_a k_a^j k_a^l \nu^i,$$
$$\text{ind}(V) = \frac{1}{6} d_{ijl} \sum_{a=1}^{5} k_a^i k_a^j k_a^l.$$

Hence, a matrix $K = (k_a^i)$ satisfies the Chern class consistency conditions (2.2) iff

$$\sum_{a=1}^{5} k_a^i \stackrel{!}{=} 0, \qquad -\frac{1}{2} d_{ijl} \sum_{a=1}^{5} k_a^j k_a^l \stackrel{!}{\leq} c_{2i}(TX), \tag{2.10}$$

for all $i = 1,\ldots,h$ and it satisfies the slope conditions (2.3) iff there exit Kähler parameters $t^i > 0$ such that

$$\mu(\mathcal{O}_X(k_a)) = d_{ijl} t^i k_a^j k_a^l \stackrel{!}{=} 0 \tag{2.11}$$

for all $a = 1,\ldots,5$. The condition (2.5) for the correct chiral asymmetry translates into

$$\frac{1}{6} d_{ijl} \sum_{a=1}^{5} k_a^i k_a^j k_a^l \stackrel{!}{=} -3|\Gamma|. \tag{2.12}$$

We have now expressed all conditions for a SLM in terms of the matrix K. The cohomology conditions (2.6) and (2.7) are not so easily dealt with. Standard methods to compute line bundle cohomology, usually based on a version of Cech cohomology, are complicated and usually algorithmic in nature. On the other hand, it has been observed in Refs. [17–20] that line bundle cohomology dimensions can be described by relatively simple formulas, which are piecewise polynomial in the line bundle integers k^i. It has also been shown in Ref. [21] that these formulas can be obtained using machine learning techniques. We will rely on some of these results in Sec. VI when we attempt to machine learn the number of Higgs multiplets.

[4]To be precise, the Hosotani mechanism that breaks the $SU(5)$ GUT group down to the standard model gauge group is available unless the discrete Wilson line takes values in $\mathbb{Z}_5$ [22] (cf. see Table I). Since the heterotic line bundle models have a reduced structure group, additional $U(1)$ gauge fields may in principle be present at low energy in case they survive the Green-Schwarz mechanism. See also the discussions in Sec. 3 of Ref. [11].

[5]One may as well imagine that one of the three $\bar{\mathbf{5}}$ multiplets leads to the down-type Higgs. In order to give a sizable mass to the top quark, however, an up-type Higgs still needs to be present, which, in turn calls for an additional down-type Higgs as well. Therefore, Eq. (2.7) should still apply.

[6]For simplicity, we assume that the Kähler cone of X is given by all $J = t^i J_i$ with $t^i \geq 0$ and the Mori cone by all positive linear combinations of the ν^i. This is indeed the case for all the CICY three-folds studied in this paper, as listed in Table I (see Ref. [26] for the comparison between the Kähler cone of a CICY and that of its ambient space).

[7]The results for the second Chern class and the index are valid provided that $c_1(V) = 0$.

III. DATASETS

In Ref. [13], complete sets of SLMs of the type described above have been found, by brute-force scanning, for all complete-intersection Calabi-Yau manifolds in products of projective spaces (CICYs) [23,24] with $h = h^{1,1}(X) \leq 6$. The data for these manifolds and the lists of matrices K for SLMs on these manifold can be found here [25]. There are 12 pairs of $(X, |\Gamma|)$ for which the set of SLMs is sufficiently large (meaning at least 1000 models) to make machine learning viable and the basic properties for these cases are listed in Table I. We emphasize that these sets contain models with and without Higgs multiplets and we will use this fact in Sec. VI. As an example, the distribution of the number of Higgs pairs (given by $h^2(X, \wedge^2 V)$) for the largest dataset in Table I, the manifold #5302 with $|\Gamma| = 4$, is shown in Fig. 1. As is evident from the plot, about a quarter of the SLMs have no Higgs pair and the remaining three quarters have one or more than one pair.

A useful way to characterize models is by the norm

$$|K| := \left(\sum_{a,i} |k_a^i|^2\right)^{1/2}, \tag{3.1}$$

which is an indicator of the size of the entries of K. As a typical example, the distribution of $|K|$ for CICY #5302 with $|\Gamma| = 4$ is shown in Fig. 2 (plot on the left).

For learning of SLMs in Sec. IV and in Sec. V we combine the lists of SLMs with random nonstandard models which are defined by random matrices K of the appropriate size. How should these random matrices be generated? One option is to generate random entries in the range observed for the SLMs (typically from -5 to 5) with a flat probability distribution. However, this method generated a distribution of $|K|$-values with a large mean and practically no overlap with that for the SLMs. Hence, random models generated in this way can be easily distinguished from SLMs by the value of $|K|$.

TABLE I. Properties of CICYs with $h^{1,1}(X) \leq 6$ and with at least 1000 standardlike rank five line bundle models. The ID of X refers to the ordering of the standard list in Refs. [23–25].

ID of X	$h = h^{1,1}(X)$	$(c_{2i}(TX))$	$\lvert\Gamma\rvert$	Number of SLMs
5256	5	(24,24,24,24,40)	4	2128
5452	5	(24,24,24,24,40)	4	2122
6890	5	(24,24,24,24,56)	2	1750
6927	5	(24,24,24,24,64)	4	1264
7487	5	(24,24,24,24,24)	4	2115
3413	6	(36,36,36,36,36,36)	3	1737
4109	6	(24,24,24,24,36,36)	2	2058
5273	6	(24,24,24,24,36,36)	2	6753
5302	6	(24,24,24,24,24,24)	2	6294
5302	6	(24,24,24,24,24,24)	4	17329
5425	6	(24,24,24,24,44,44)	2	3128
6738	6	(24,24,24,24,44,44)	2	4243

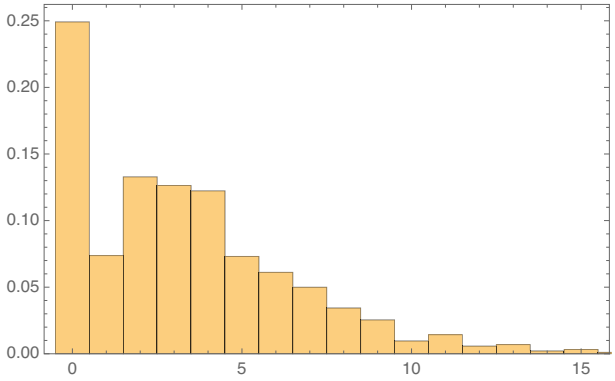

FIG. 1. Fractional distribution of the number of Higgs pairs for SLMs on the manifold #5302 with $|\Gamma| = 4$.

Instead, we generate entries with a normal distribution, where the mean and width are taken the same as observed for the SLMs. This ensures that the matrices K for SLMs and non-SLMs have a similar distribution of entries. For all datasets in Table I, the mean of the entries is close to 0 and the standard deviation of the entries is between 1 and 1.3, depending on the dataset. In addition, we ensure that all random models satisfy the trivial Chern class condition $c_1(V) = 0$, that is, the first Eq. (2.10). For CICY #5302 and $|\Gamma| = 4$ the combined distribution of SLMs and non-SLMs is shown on the right-hand side of Fig. 2.

What is the distribution of the properties which define SLMs within this set of random non-SLMs, generated as described above? The anomaly condition, that is, the second Eq. (2.10), is satisfied by practically all random models (more precisely, by a fraction of 0.94–0.99 depending on the manifold). This is expected, given we have chosen random matrices with entries similar to those for the SLMs and the condition in question is an inequality. On the other hand, the slope condition (2.11) is almost never satisfied for the random models (at most for a fraction of 0.01). Finally, the random models also very rarely satisfy the family condition (2.12) (a fraction of 0.02 to 0.06, depending on the manifold). A neural network which distinguishes SLMs from random non-SLMs will have to be sensitive to at least one of the last two conditions. A network with a success rate close to 1 needs to be sensitive to each of the conditions separately.

For the practical machine learning application, we will focus on a particular manifold X and symmetry order $|\Gamma|$, corresponding to one of the rows in Table I. For supervised learning of the standard model property on this manifold we will use a dataset of the form

$$\{K = (k_a^i) \to 0 \text{ or } 1\}, \tag{3.2}$$

where K is either a random integer matrix, generated as explained above, describing a non-SLM if assigned to 0 or a matrix describing a SLM if assigned to 1.

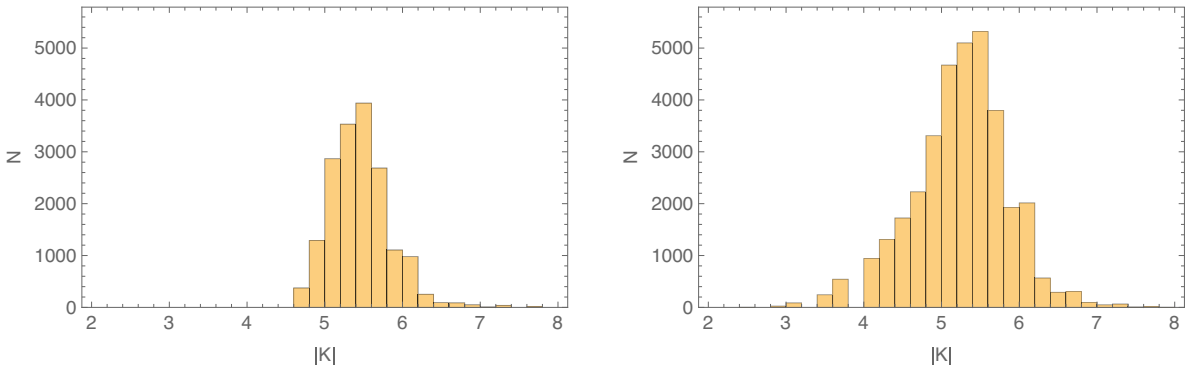

FIG. 2. Distribution of $|K|$ in Eq. (3.1) for SLMs (left) and for SLMs and non-SLMs combined (right) for CICY #5302 and $|\Gamma| = 4$.

For unsupervised learning of the standard model property we use the same dataset $\{K\}$ but with the labels omitted.

To learn about the Higgs, the datasets will be of the form

$$\{K = (k_a^i) \to 0 \text{ or } 1\}$$
$$\{K = (k_a^i) \to \#\text{ of Higgs pairs}\}, \tag{3.3}$$

where we only include matrices K which describe SLMs. Recall that SLMs can have varying numbers of Higgs pairs, as shown in Fig. 1. The first above set is appropriate for a simple binary classification of the absence (0) or presence (1) of Higgs multiplets and the second set is used to learn the actual number of Higgs multiplets.

As we will see in Sec. VI, training sets of this form are not quite suited for successful learning of the Higgs property. Instead, we will be using the feature-enhanced datasets

$$\{(k_a^i, k_a^i k_a^j, k_a^i k_a^j k_a^l) \to 0 \text{ or } 1\}$$
$$\{(k_a^i, k_a^i k_a^j, k_a^i k_a^j k_a^l) \to \#\text{ of Higgs pairs}\}, \tag{3.4}$$

with the quadratic and cubic monomials in the line bundle integers k_a^i added to the input. This method is informed by the observation [17–21] that line bundle cohomology dimensions on three-folds are described by piecewise cubic formulas.

For the purpose of Higgs learning, we will also find that our datasets as in Table I are still too small. In this context, it is useful to observe that our models $K = (k_1, \ldots, k_5)$ have an obvious permutation symmetry S_5 which permutes the five line bundles k_a. Model properties are of course completely independent of these permutations and the resulting redundancies have already been removed from the datasets in Table I. Conversely, we can now use these permutation to enhance the size of our datasets. We will return to this point in Sec. VI.

Throughout the paper, neural networks will be realized both in TENSORFLOW and in *Mathematica*, and trained with a standard stochastic gradient descent method and using a mean square loss. Training and validation will be measured by the loss and by the success rate, by which we mean the fraction of models for which the trained network produces the correct integer target after rounding.

IV. LEARNING STANDARD MODELS WITH SUPERVISED LEARNING

In this section, we will study whether neural networks can distinguish SLMs from non-SLMs. Our datasets are of the form (3.2) and are split into a training set (70%) and a validation set (30%). As we will see, for the purposes of this section, it is sufficient to consider relatively simple networks of the form

$$K \in \mathbb{Z}^{5h} \xrightarrow{L} \mathbb{R}^{16} \xrightarrow{\text{selu}} \mathbb{R}^{16} \xrightarrow{L} \mathbb{R} \xrightarrow{\sigma} [0,1] \subset \mathbb{R}, \tag{4.1}$$

where L is an affine transformation $x \mapsto Wx + b$ with a weight matrix W and a bias vector b of the appropriate dimensions, selu is the scaled exponential linear unit activation function, defined for each element as

$$\text{selu}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha e^x - \alpha, & \text{if } x \le 0, \end{cases} \tag{4.2}$$

and σ is the logistic sigmoid activation function

$$\sigma(x) = (1 + e^{-x})^{-1}. \tag{4.3}$$

We train this network for all cases listed in Table I and find that the training and validation success rates are always ≥ 0.99. We conclude that, for a given compactification space, neural networks of the form (4.1) can successfully distinguish SLMs from non-SLMs. This works for a range of compactification spaces, as in Table I.

The successful generalization of the network (4.1) for the problem at hand motivates tackling a somewhat more ambitious task. The main problem in searching for SLMs systematically is the shear number of integer matrices K (with $c_1(V) = 0$) which increases with a power $4h$. For $h \ge 7$ a full systematic search basically becomes impossible but it is still feasible to scan matrices K with entries

constrained in a suitably small range. Suppose we generate a training set of matrices with small entries from such a restricted scan. Does a network trained on such a set generalize to matrices with larger entries?

In order to answer this question, we focus on the manifold #5302 with $|\Gamma| = 4$ from Table I, which provides our largest dataset. We select from this set all matrices with $|K| \leq 5$ (roughly 1/3 of the models, corresponding to the left part of the distribution in Fig. 2), which we split into a training set (70%) and a validation set (30%). The other matrices with $|K| > 5$ (roughly 2/3 of the models, corresponding to the right part of the distribution in Fig. 2) are used as a test set in order to test the generalization of the network to matrices with a size beyond the training range.

For the data structured in this way, we train the network (4.1). The training and validation loss as a function of training rounds is shown in Fig. 3 and training and validation success rates are both ≥ 0.99. Crucially, the success rate of the network trained on this set of matrices with $|K| \leq 5$ on the test set of matrices with $|K| > 5$ is $\geq$0.98, and the values of Matthew's phi and F-score are 0.991 and 0.9956, respectively. In other words, the network generalizes extremely well to matrices with larger entries.

For this case, it is instructive to look at the three properties which are required for a SLM, namely the anomaly condition [second Eq. (2.10)], the slope condition (2.11) and the family condition (2.12), separately. Table II presents the relevant confusion matrix, covering all eight combinations of these three properties, for the complete dataset including all values of $|K|$. As discussed earlier, the eight combinations of properties have very different frequencies. However, the message from Table II is that the network does well to distinguish SLMs from non-SLMs for most combinations of properties. (The combination (0,1,1) of slope zero models with the right family number which fail on the anomaly condition has such a low frequency that the success rates are not conclusive.)

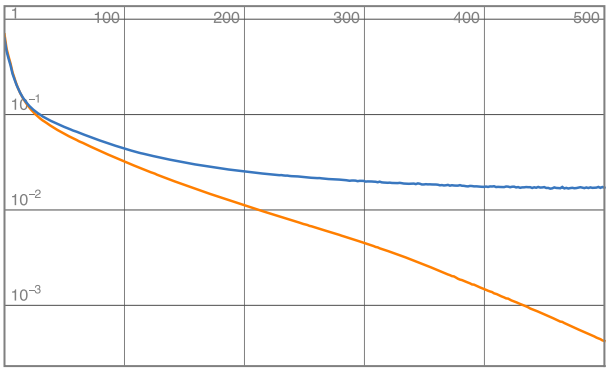

FIG. 3. Training loss (orange) and validation loss (blue) for the network (4.1) as a function of training rounds for the data from CICY #5302 and $|\Gamma| = 4$ with $|K| \leq 5$.

TABLE II. Confusion matrix for the dataset on the CICY #5302 with $|\Gamma| = 4$ and the network trained on models with $|K| \leq 5$. The left column indicates the various combinations of properties found in the dataset and the final two columns give the fractions of those models identified as SLMs and non-SLMs.

(Anomaly, slope, family)	Number of models	Non-SMs (0)	SMs (1)
(0, 0, 0)	573	0.96	0.04
(0, 0, 1)	33	0.91	0.09
(0, 1, 0)	46	0.96	0.04
(0, 1, 1)	2	1.00	0.00
(1, 0, 0)	15956	0.99	0.01
(1, 0, 1)	517	1.00	0.00
(1, 1, 0)	197	0.98	0.02
(1, 1, 1)	17329	0.02	0.98

V. AUTOENCODING STANDARD MODELS

We have seen that supervised learning can be used to distinguish SLMs from non-SLMs quite efficiently. In this section, we will investigate whether something similar can be accomplished in the context of unsupervised learning, using an autoencoder.

We will focus on the manifold #5302 with $|\Gamma| = 4$, our largest training set and, as before, we split into subsets with $|K| \leq 5$, used for training, and with $|K| > 5$, used for testing. Guided by the observation in Ref. [14] we also one-hot encode the matrices K which leads to binary input vectors K_{hot} of dimension 330. The structure of the encoder is

$$K_{\text{hot}} \in \mathbb{Z}^{330} \xrightarrow{\tilde{L}} \mathbb{R}^{32} \xrightarrow{\tilde{L}} \mathbb{R}^{16} \xrightarrow{\tilde{L}} \mathbb{R}^{8} \xrightarrow{\tilde{L}} \mathbb{R}^{2}, \tag{5.1}$$

while the decoder

$$\mathbb{R}^{2} \xrightarrow{\tilde{L}} \mathbb{R}^{8} \xrightarrow{\tilde{L}} \mathbb{R}^{16} \xrightarrow{\tilde{L}} \mathbb{R}^{32} \xrightarrow{\tilde{L}} \mathbb{Z}^{330} \ni \hat{K}_{\text{hot}}, \tag{5.2}$$

maps between the same dimensions, in reverse order. Here, $\tilde{L} = \text{selu} \circ L$ is the combination of an affine transformation $x \mapsto Wx + b$ between the appropriate dimensions and a selu activation function (4.2). Training is performed by minimising the mean square difference $|K_{\text{hot}} - \hat{K}_{\text{hot}}|$, as usual. Note that the latent space (the output space of the encoder) is $\mathbb{R}^2$, so the compression of the matrix K facilitated by this autoencoder is to two dimensions.

We remark that one could perform analyses using other methods of dimensional reduction and interestingly arrive at similar results to Fig. 4. For instance, a principal component analysis was carried out to reduce the initial labeled data (without hot-encoding but just flattening the bundles into an integer vector of length $6 \times 5 = 30$) directly to 2 dimensions. One finds that for each of the $|K| \leq 5$ and $|K| > 5$ cases, there is a separation between the SLMs and non-SLMs much like Fig. 4.

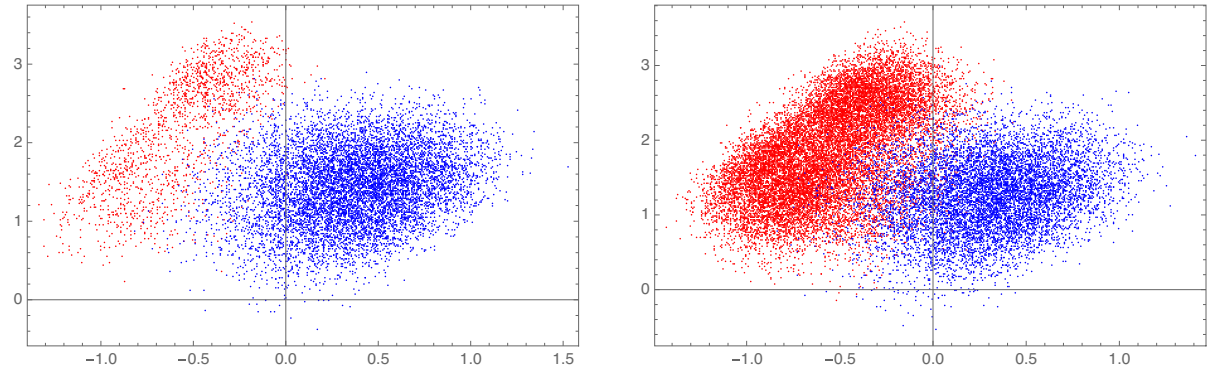

FIG. 4. The models on manifold #5302 with $|\Gamma| = 4$ encoded in the two-dimensional latent space of the autoencoder (5.1), (5.2), trained on models with $|K| \leq 5$. Red points represent SLMs and blue points non-SLMs. The plot on the left (right) is for models with $|K| \leq 5$ ($|K| > 5$).

From the above autoencoder, trained with the subset of matrices with $|K| \leq 5$, we then extract the encoder (5.1) and compute the image of all models in the two-dimensional latent space. The results are shown in Fig. 4. As is evident from those plots, the separation between SLMs and non-SLMs in the two-dimensional latent space is quite convincing. We emphasize that the autoencoder was trained on the models with $|K| \leq 5$ only while the models with $|K| > 5$ are "unseen data." Nevertheless, the separation between SLMs and non-SLMs occurs for the training set (left plot in Fig. 4) as well as for the test set (right plot in Fig. 4). Just as in the case of supervised learning, the autoencoder, therefore, generalizes beyond the training range.

VI. LEARNING ABOUT HIGGS MULTIPLETS

So far we have used neural networks to distinguish SLMs and non-SLMs. The properties separating these two classes of models are topological in nature so we have studied machine learning of topological quantities. In this section, we will be more ambitious and try to learn properties which are nontopological. Specifically, we will restrict our datasets to SLMs and attempt to learn the presence/absence of Higgs multiplets and, in a second step, the number of Higgs multiplets.

More specifically, we will focus on our largest dataset of SLMs on the manifold #5302 with $|\Gamma| = 4$ with 17329 models. The distribution of the number of Higgs pairs for this dataset is shown in Fig. 1. Roughly a quarter of the models have no Higgs pairs and all others have one or more than one Higgs pair.

We begin with the less ambitious task of learning the presence or absence of Higgs pairs, so the structure of the dataset is $\{K \to 0 \text{ or } 1\}$ as in Eq. (3.3). As usual we use 70% of these models for training and the other 30% for validation. We consider a variety of networks of the form

$$K \in \mathbb{Z}^n \xrightarrow{\tilde{L}} \mathbb{R}^{n_1} \xrightarrow{\tilde{L}} \cdots \xrightarrow{\tilde{L}} \mathbb{R}^{n_l} \xrightarrow{L} \mathbb{R} \xrightarrow{\sigma} [0,1], \qquad (6.1)$$

where $\tilde{L} = \text{selu} \circ L$ combines an affine transformation L with a selu activation (4.2) and σ is a logistic sigmoid activation (4.3). We vary the widths n_i as well as the depth l of the network. A typical result for training and validation loss for the choice $l = 4$, $n = 30$ and $(n_1, n_2, n_3, n_4) = (256, 128, 64, 16)$ is shown in Fig. 5. The confusion matrices for this case are given by

$$\begin{array}{c} \\ \text{target 0} \\ \text{target 1} \end{array} \begin{array}{c} \text{training set} \\ \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0.01 & 0.23 \\ 0.01 & 0.75 \end{pmatrix} \end{array} \quad \begin{array}{c} \text{validation set} \\ \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0.01 & 0.25 \\ 0.01 & 0.73 \end{pmatrix} \end{array} . \qquad (6.2)$$

Both the loss plot and the confusion matrices indicate complete failure of generalization. In fact, the confusion matrices shows that all models, including the quarter without Higgs pairs, is identified by the network as having Higgs pairs. We have considered the network (6.1) for a variety of depths and widths and with a number of standard measures to improve generalization, including optimizing hyperparameters, different activation functions, inclusion of batch normalization and dropout layers. The results remain essentially unchanged from the above example. It appears a simple network (6.1) is not capable of learning a nontopological property, such as the presence or absence of Higgs pairs.

What saves the day is an insight into the structure of line bundle cohomology. It is conjectured and, in some cases, proved [17–21], that line bundle cohomology dimensions on three-folds are described by piecewise cubic polynomials in the line bundle integers k^i. As we have discussed, the number of Higgs pairs is governed by line bundle cohomology, so this insight is relevant. It suggests we should feature-enhance our datasets and add to the line bundle integers $K = (k_a^i)$ their quadrics $(k_a^i k_a^j)$ and their cubics $(k_a^i k_a^j k_a^l)$ (for $i \leq j \leq l$). This means we now consider a dataset of the form $\{(k_a^i, k_a^i k_a^j, k_a^i k_a^j k_a^l) \to 0 \text{ or } 1\}$, as in Eq. (3.4).

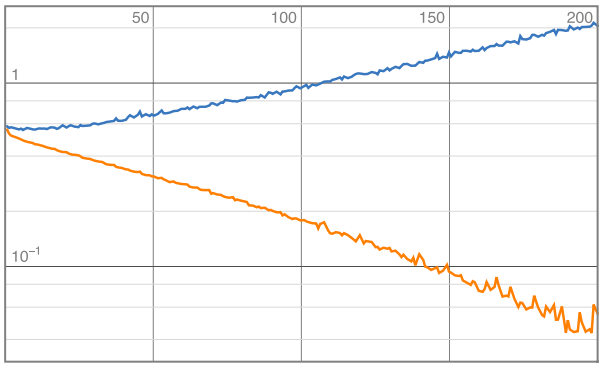

FIG. 5. Training loss (orange) and validation loss (blue) for the network (6.1) with $l = 4$, $n = 30$ and $(n_1, n_2, n_3, n_4) = (256, 128, 64, 16)$, attempting to learn the absence/presence of Higgs pairs for SLMs on the manifold #5032 with $|\Gamma| = 4$. The increase of the validation loss indicates a failure to generalize.

However, there is one further problem. The dimensions of the enhanced feature space is now 415 and the available training set of 17329 models is simply not large enough to train efficiently for such a high dimension. We can solve this problem by using the S_5 symmetry which permutes the five line bundles. For each of our models we generate 19 random permutations and thereby increase the size of our dataset by a factor of 20. As usual, we split this set into a training set (70%) and a validation set (30%) and we use a network (6.1) with $l = 3$, $n = 415$ and $(n_1, n_2, n_3) = (256, 64, 16)$. The result for training and validation loss is in Fig. 6, which shows that both training and validation are successful.

The confusion matrix

$$\begin{array}{c} \\ \text{target } 0 \\ \text{target } 1 \end{array} \begin{array}{c} \begin{array}{cc} \text{trainingset} & \\ 0 & 1 \end{array} \\ \begin{pmatrix} 0.25 & 0 \\ 0 & 0.75 \end{pmatrix} \end{array} \quad \begin{array}{c} \begin{array}{cc} \text{validation set} & \\ 0 & 1 \end{array} \\ \begin{pmatrix} 0.25 & 0 \\ 0 & 0.75 \end{pmatrix} \end{array} \tag{6.3}$$

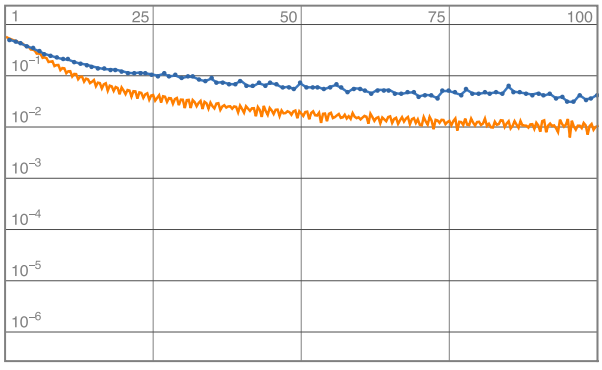

FIG. 6. Training loss (orange) and validation loss (blue) for the network (6.1) with $l = 3$, $n = 415$ and $(n_1, n_2, n_3) = (256, 64, 16)$, learning the absence/presence of Higgs pairs for SLMs on the manifold #5032 with $|\Gamma| = 4$. A feature-enhanced dataset (3.4) is used.

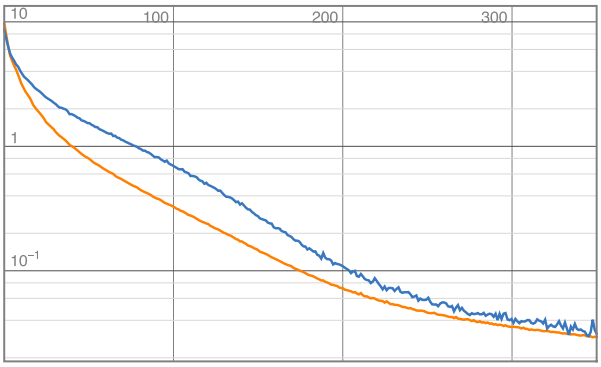

FIG. 7. Training loss (orange) and validation loss (blue) for the network (6.4) with $l = 3$, $n = 415$ and $(n_1, n_2, n_3) = (256, 64, 16)$, learning the number of Higgs pairs for SLMs on the manifold #5032 with $|\Gamma| = 4$. A feature-enhanced dataset (3.4) is used.

indicates models with and without Higgs pairs are recognized with success rate 1.

Encouraged by this success we now attempt to learn the number of Higgs multiplets for manifold #5302 with $|\Gamma| = 4$, using a feature-enhanced dataset $\{(k_a^i, k_a^i k_a^j, k_a^i k_a^j k_a^l) \to \#\text{Higgs pairs}\}$, as in Eq. (3.4). As before, we increase the size of the dataset 20-fold by using the invariance under permutations of the five line bundles. Networks are similar to (6.1) but with the final logistic sigmoid activation omitted, as appropriate for more general integer target values:

$$K \in \mathbb{Z}^n \xrightarrow{\tilde{L}} \mathbb{R}^{n_1} \xrightarrow{\tilde{L}} \cdots \xrightarrow{\tilde{L}} \mathbb{R}^{n_l} \xrightarrow{L} \mathbb{R}. \tag{6.4}$$

A network of this type with depth $l = 3$, $n = 415$ and $(n_1, n_2, n_3) = (256, 64, 16)$ trains and validates successfully, as shown in Fig. 7. The confusion matrices for both training and validation sets are nearly diagonal, with all off-diagonal entries < 0.01, and the diagonal entries correctly reproduce the fractional distribution in Fig. 1, with deviations uniformly less than 0.01.

VII. CONCLUSION

In this paper, we have demonstrated that neural networks are capable of learning phenomenologically relevant properties of string compactifications. Our datasets have been constructed from heterotic line bundle models [11–13], one of the few datasets where considerable numbers of models with the standard model spectrum have been constructed.

For a given Calabi-Yau three-fold X, a line bundle model is defined by a sum $V = \oplus_{a=1}^5 \mathcal{O}_X(k_a)$ of five line bundles, where k_a are $h = h^{1,1}(X)$ dimensional integer vectors classifying line bundles. Hence, such models can be represented by a $h \times 5$ integer matrices $K = (k_a^i)$ which constitute the features in our applications to machine learning. In Ref. [13] complete intersection Calabi-Yau manifolds with $h \leq 6$ have been scanned for phenomenologically interesting

line bundle models. Sizeable datasets (see Table I) of standardlike models (SLMs), that is consistent models with the correct gauge group and the correct chiral asymmetry, have been found. These datasets, described in Table I, form the basis of the machine learning applications we have discussed. Many further properties of these models have been computed [12] and we will rely on the results for the number of Higgs pairs.

Our first task has been to learn topological properties of line bundle models. To this end, we have merged SLMs for a given Calabi-Yau manifold with similar sized datasets of random models to obtain sets of the form $\{K \to 0 \text{ or } 1\}$, where 1 indicates an SLM and 0 a random model. Training is carried out one manifold at a time. Relatively simple networks of the form (4.1) lead to near perfect success rates of ≥ 0.99 on both training and validation sets for all cases in Table I. In a more ambitious approach, for the manifold #5302 with symmetry order $|\Gamma| = 4$ (see Table I), we have used a training set drawn from matrices K with small entries ($|K| \leq 5$) and a test set of matrices with large entries ($|K| > 5$). It turns out that neural networks (4.1) trained on such a training set still perform extremely well, at a success rate of ≥ 0.98, on the test set. In other words, the neural network generalizes well beyond the "training range."

For the same dataset, on manifold #5302, split into a training set with small entries and a test set with large entries, we have considered unsupervised learning with an autoencoder (5.1), (5.2). The results, summarized in Figure 4, show that the autoencoder is capable of distinguishing SLMs from non-SLMs and that it retains this capacity on the test set. In particular, this means the generalization beyond the training range observed for supervised learning, persists for the autoencoder.

The final task has been to learn a nontopological property, namely the absence or presence (or, more ambitiously, the number) of Higgs pairs. This property is determined by the values of line bundle cohomology dimensions. For this task we have focused on the dataset of SLMs on manifold #5302 with symmetry order $|\Gamma| = 4$ which provides the largest number of models (about 17000). We have started with simple datasets $\{K \to 0 \text{ or } 1\}$, where 1 is for the presence and 0 for the absence of Higgs pairs, and fully connected networks of the form (6.1). This simple approach resulted in complete failure to generalize, as exemplified in Fig. 5. Inspired by the observation [17–21] that line bundle cohomology dimensions on three-folds are described by piecewise cubic expressions in the line bundle integers k_a^i we have feature-enhanced our dataset to $\{(k_a^i, k_a^i k_a^j, k_a^i k_a^j k_a^l) \to 0 \text{ or } 1\}$, that is, by including quadratic and cubic monomials. Given the significantly increased dimension of the feature space (to 415) we have also increased the size of the dataset by using the invariance under permutations of the five line bundles. A simple fully connected neural network (6.1) trained with this enhanced dataset does indeed succeed and distinguishes models with and without Higgs pairs reliably. We have also trained a network (6.4) with a dataset $\{(k_a^i, k_a^i k_a^j, k_a^i k_a^j k_a^l) \to \#\text{ of Higgs pairs}\}$ to predict the number of Higgs multiplets successfully.

Our results show that neural networks are capable of learning both topological and nontopological properties of string vacua and are able to select models with phenomenological promise. Recognizing nontopological properties is considerably more difficult and has been achieved by feature engineering based on insight into the underlying mathematical structure. The observed generalization beyond the training range is encouraging. It implies that networks trained on models with small flux integers—which may be obtained by systematic scanning—can be used to identify promising models for larger flux integers. Our results may also be helpful in designing more complicated networks, required for generative approaches (GANs) or reinforcement learning, for instance.

It would be interesting to study whether the present results for heterotic line bundle models persist for other classes of string models, for example for F-theory models or for heterotic models with non-Abelian bundles.

ACKNOWLEDGMENTS

R. D. is supported by a Skynner Fellowship of Balliol College, Oxford. Y. H. H. acknowledges Science and Technology Facilities Council UK, for grant No. ST/J00037X/1. The work of S. J. L. is supported by IBS under the project code, IBS-R018-D1. A. L. gratefully acknowledges support from the Simons Center for Geometry and Physics, Stony Brook University at which some of the research for this paper was performed.

[1] Y. H. He, Deep-learning the landscape, arXiv:1706.02714.
[2] Y. H. He, Machine-learning the string landscape, Phys. Lett. B **774** (2017) 564.
[3] D. Krefl and R. K. Seong, Machine learning of Calabi-Yau volumes, Phys. Rev. D **96**, 066014 (2017).
[4] F. Ruehle, Evolving neural networks with genetic algorithms to study the string landscape, J. High Energy Phys. 08 (2017) 038.
[5] J. Carifio, J. Halverson, D. Krioukov, and B. D. Nelson, Machine learning in the string landscape, J. High Energy Phys. 09 (2017) 157.

[6] F. Ruehle, Data science applications to string theory, Phys. Rep. **839**, 1 (2020).
[7] Y. H. He, The Calabi-Yau landscape: From geometry, to physics, to machine-learning, arXiv:1812.02893.
[8] W. Taylor and Y. N. Wang, The F-theory geometry with most flux vacua, J. High Energy Phys. 12 (2015) 164.
[9] T. Hübsch (private communication).
[10] A. Constantin, Y. H. He, and A. Lukas, Counting string theory standard models, Phys. Lett. B **792**, 258 (2019).
[11] L. B. Anderson, J. Gray, A. Lukas, and E. Palti, Two hundred heterotic Standard Models on smooth Calabi-Yau threefolds, Phys. Rev. D **84** (2011) 106005.
[12] L. B. Anderson, J. Gray, A. Lukas, and E. Palti, Heterotic line bundle standard models, J. High Energy Phys. 06 (2012) 113.
[13] L. B. Anderson, A. Constantin, J. Gray, A. Lukas, and E. Palti, A comprehensive scan for heterotic SU(5) GUT models, J. High Energy Phys. 01 (2014) 047.
[14] A. Mutter, E. Parr, and P. K. S. Vaudrevange, Deep learning in the heterotic orbifold landscape, Nucl. Phys. **B940** (2019) 113.
[15] J. Halverson, B. Nelson, and F. Ruehle, Branes with brains: Exploring string vacua with deep reinforcement learning, J. High Energy Phys. 06 (**2019**) 003.
[16] M. Larfors and R. Schneider, Explore and exploit with heterotic line bundle models, Fortschr. Phys. **68**, 2000034 (2020).
[17] A. Constantin and A. Lukas, Formulae for line bundle cohomology on Calabi-Yau threefolds, Fortschr. Phys. **67**, 1900084 (2019).
[18] D. Klaewer and L. Schlechter, Machine learning line bundle cohomologies of hypersurfaces in toric varieties, Phys. Lett. B **789** (2019) 438.
[19] M. Larfors and R. Schneider, Line bundle cohomologies on CICYs with Picard number two, Fortschr. Phys. **67**, 1900083 (2019).
[20] C. R. Brodie, A. Constantin, R. Deen, and A. Lukas, Index formulae for line bundle cohomology on complex surfaces, Fortschr. Phys. **68**, 1900086 (2020).
[21] C. R. Brodie, A. Constantin, R. Deen, and A. Lukas, Machine learning line bundle cohomology, Fortschr. Phys. **68**, 1900087 (2020).
[22] B. McInnes, Group theoretic aspects of the Hosotani mechanism, J. Phys. A **22**, 2309 (1989).
[23] P. Candelas, A. M. Dale, C. A. Lutken, and R. Schimmrigk, Complete intersection Calabi-Yau manifolds, Nucl. Phys. **B298**, 493 (1988).
[24] P. Green and T. Hubsch, Calabi-Yau manifolds as complete intersections in products of complex projective spaces, Commun. Math. Phys. **109**, 99 (1987).
[25] Heterotic line bundle data can be downloaded at http://www-thphys.physics.ox.ac.uk/projects/CalabiYau/linebundlemodels/index.html.
[26] L. B. Anderson, X. Gao, J. Gray, and S. J. Lee, Fibrations in CICY threefolds, J. High Energy Phys. 10 (2017) 077.