Available online at www.sciencedirect.com

ScienceDirect

Nuclear Physics B 952 (2020) 114922

NUCLEAR PHYSICS B

www.elsevier.com/locate/nuclphysb

Contrast data mining for the MSSM from strings

Erik Parr *, Patrick K.S. Vaudrevange

Physik Department T75, Technische Universität München, James–Franck–Straße 1, 85748 Garching, Germany

Received 7 November 2019; received in revised form 14 December 2019; accepted 8 January 2020
Available online 13 January 2020
Editor: Stephan Stieberger

Abstract

We apply techniques from data mining to the heterotic orbifold landscape in order to identify new MSSM-like string models. To do so, so-called contrast patterns are uncovered that help to distinguish between areas in the landscape that contain MSSM-like models and the rest of the landscape. First, we develop these patterns in the well-known $\mathbb{Z}_6$-II orbifold geometry and then we generalize them to all other $\mathbb{Z}_N$ orbifold geometries. Our contrast patterns have a clear physical interpretation and are easy to check for a given string model. Hence, they can be used to scale down the potentially interesting area in the landscape, which significantly enhances the search for MSSM-like models. Thus, by deploying the knowledge gain from contrast mining into a new search algorithm we create many novel MSSM-like models, especially in corners of the landscape that were hardly accessible by the conventional search algorithm, for example, MSSM-like $\mathbb{Z}_6$-II models with $\Delta(54)$ flavor symmetry.

1. Introduction

String theory is a promising candidate for a UV-complete theory of quantum gravity. However, to prove its usefulness it has to incorporate the standard model (SM) or its supersymmetric extension (the MSSM) as a low-energy limit. Thus, one of the main tasks of string phenomenology is to find a string model that is consistent with all experimental and observational data – or, at least, that comes close to the MSSM, i.e. a model that is MSSM-like. This task is very difficult,

* Corresponding author.
E-mail addresses: erik.parr@tum.de (E. Parr), patrick.vaudrevange@tum.de (P.K.S. Vaudrevange).

https://doi.org/10.1016/j.nuclphysb.2020.114922

mainly due to two reasons: (i) String theory predicts extra spatial dimensions. Hence, the connection between string theory and the MSSM must be related to the way how the extra dimensions are compactified. However, the number of different compactifications is huge [1,2], giving rise to the so-called string landscape of effective 4D theories originating from string theory. (ii) String theory is very predictive because after specifying the compactification the effective 4D theory, including all symmetries, particles and couplings, is completely fixed.

In the case of the ten-dimensional $E_8 \times E_8$ heterotic string we have to compactify six dimensions. To do so, we choose six-dimensional toroidal orbifolds [3,4] as they allow for a world-sheet formulation of string theory instead of a supergravity approximation, see e.g. [5–15] for other schemes. Then, the conventional approach to search for MSSM-like string models from heterotic orbifolds is given by a random scan in the landscape [16–18] or in fertile islands that can be identified by certain patterns, like local GUTs [19–21]. In this paper, we show that the approach of defining fertile islands can be generalized by applying machine learning techniques to the string landscape [22–25], see also [26–33]. A first hint towards such a generalization was observed in ref. [34]: a neural network was able to identify additional patterns and to cluster the models of the heterotic orbifold landscape according to them. Surprisingly, MSSM-like models turned out to be localized close to each other on fertile islands, even though the neural network did not know whether a given model is MSSM-like or not, denoted by ~~MSSM~~-like.

In this paper we propose and demonstrate a new search strategy for MSSM-like models based on additional patterns that is superior by orders of magnitude. This search is developed from the knowledge gained by analyzing the heterotic orbifold landscape with tools from data mining. Data mining has been developed for the purpose to prepare, visualize and explore huge databases. Hence, the suitability of data mining to the string landscape is obvious. In particular, we apply a special setup called contrast data mining to the heterotic orbifold landscape. The basic idea of contrast data mining is to identify so-called contrast patterns that allow us to focus our search on those areas in the landscape where the MSSM-like models reside. Our contrast patterns have a clear physical interpretation: they are given by the number of unbroken roots in the hidden E_8 factor [9,35] and the numbers of various bulk matter fields.

This paper is structured as follows: In section 2 we review the conventional search algorithm for MSSM-like models in the heterotic orbifold landscape. Then, we improve this algorithm using traditional methods: first, we take the Weyl symmetry into account and, secondly, in section 3 we include phenomenological constraints. These improvements already reduce the number of models that have to be scanned by the search algorithm in the test case of the $\mathbb{Z}_6$-II orbifold geometry. Afterwards, in section 4 we apply data mining techniques to our $\mathbb{Z}_6$-II dataset. Doing so, we can identify contrast patterns that help to distinguish between areas in the landscape that contain MSSM-like models and others. We implement these contrast patterns into our search algorithm and show that we can easily construct many new MSSM-like $\mathbb{Z}_6$-II models – a fact that might be surprising given the extensive searches done especially in this orbifold geometry. Remarkably, we can identify two MSSM-like $\mathbb{Z}_6$-II models with $\Delta(54)$ flavor symmetry (related to a vanishing Wilson line of order three), see section 5. Thereafter, our contrast patterns are successfully extended to all $\mathbb{Z}_N$ orbifold geometries in section 6, where Table 8 summarizes our results, see also [36]. Finally, in section 7 we conclude.

2. Searching the heterotic orbifold landscape

An orbifold compactification [3,4] is specified by a six-dimensional orbifold geometry $\mathbb{O}$ and a gauge embedding that acts on the sixteen gauge degrees of freedom of the heterotic string.

Table 1
Table of geometrical constraints for shift vectors and Wilson lines in the case of the $\mathbb{Z}_6$-II orbifold geometry.

Vector	Order N_k	Additional constraint
V_1	6	
V_2	1	not present, i.e. $V_2 = (0^{16})$
W_1	1	$W_1 = (0^{16})$
W_2	1	$W_2 = (0^{16})$
W_3	3	$W_3 = W_4$
W_4	3	
W_5	2	
W_6	2	

While there exist only 138 orbifold geometries with Abelian point group (i.e. $\mathbb{Z}_{N_1}$ or $\mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2}$) and $\mathcal{N} = 1$ supersymmetry [37], the true size of the heterotic orbifold landscape unfolds if we take the number of inequivalent gauge embeddings into account. For a general $\mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2}$ orbifold, a gauge embedding is given by two shift vectors, V_1 and V_2, and six Wilson lines, W_1 to W_6, corresponding to the six compactified directions [38]. In this paper we concentrate on the $\mathrm{E}_8 \times \mathrm{E}_8$ heterotic string,[1] which implies that each vector can be split into two eight-dimensional vectors. For example, the sixteen-dimensional shift vector V_1 consists of the eight-dimensional vectors $V_1^{(1)}$ and $V_1^{(2)}$, which act on the first and second E_8 factor, respectively. Altogether a gauge embedding is determined by a gauge embedding matrix

$$M = \begin{pmatrix} V_1^{(1)} & V_1^{(2)} \\ V_2^{(1)} & V_2^{(2)} \\ W_1^{(1)} & W_1^{(2)} \\ W_2^{(1)} & W_2^{(2)} \\ W_3^{(1)} & W_3^{(2)} \\ W_4^{(1)} & W_4^{(2)} \\ W_5^{(1)} & W_5^{(2)} \\ W_6^{(1)} & W_6^{(2)} \end{pmatrix}, \tag{1}$$

where we denote the vector in the k-th line by M_k for $k = 1, \ldots, 8$ and split it into two parts $M_k^{(\alpha)}$ for $\alpha = 1, 2$ corresponding to the two E_8 factors, for example, $M_3 = W_1 = (W_1^{(1)}, W_1^{(2)})$ for the first Wilson line W_1. Depending on the orbifold geometry, shift vectors and Wilson lines are subject to geometrical constraints that, for example, fix the order of the shift vector V_1 to N for a $\mathbb{Z}_N$ orbifold geometry, see e.g. ref. [39]. To be more specific and as we will mainly work with the so-called $\mathbb{Z}_6$-II (1, 1) orbifold geometry (using the nomenclature from ref. [37], abbreviated as $\mathbb{Z}_6$-II in the following), we summarize the geometrical constraints on the shift vectors and Wilson lines for the $\mathbb{Z}_6$-II orbifold geometry in Table 1.

[1] The methods described in this paper are easy to generalize to the SO(32) heterotic string.

In general, there are two ways to expand a sixteen-dimensional shift vector or Wilson line naturally: either in terms of the simple roots α_I of $\mathrm{E}_8 \times \mathrm{E}_8$ or in terms of the dual simple roots α_I^*, $I = 1, \ldots, 16$, where $\alpha_I^* \cdot \alpha_J = \delta_{IJ}$. Both choices give a basis of the (self-dual) root lattice $\Lambda_{\mathrm{E}_8 \times \mathrm{E}_8}$ of $\mathrm{E}_8 \times \mathrm{E}_8$. For later convenience (see section 2.3) we decide to expand the vectors in terms of the dual basis, i.e. we parameterize the vectors M_k in the gauge embedding matrix M as

$$M_k \;=\; \frac{1}{N_k} \sum_{I=1}^{16} d_{k\,I}\, \alpha_I^* \,. \tag{2}$$

Here, N_k defines the order of the shift vector or Wilson line and $d_{k\,I} \in \mathbb{Z}$ for $k = 1, \ldots, 8$ and $I = 1, \ldots, 16$ are integers. Consequently, a gauge embedding matrix M corresponds to a point in $d \in \mathbb{Z}^{128}$ since $d_{k\,I}$ has $8 \times 16 = 128$ components. Note that this construction eq. (2) inherently ensures the correct order of the respective vector. In detail, for all vectors $k = 1, \ldots, 8$ we have

$$N_k\, M_k \;\in\; \Lambda_{\mathrm{E}_8 \times \mathrm{E}_8} \,. \tag{3}$$

2.1. Conditions from modular invariance

In order to obtain consistent string compactifications we have to impose conditions from modular invariance of the one-loop string partition function on the gauge embedding matrix M. These conditions read [40]

$$N_1 \left(V_1^2 - v_1^2\right) = 0 \mod 2 \,, \tag{4a}$$

$$N_2 \left(V_2^2 - v_2^2\right) = 0 \mod 2 \,, \tag{4b}$$

$$\gcd(N_1, N_2)\ (V_1 \cdot V_2 - v_1 \cdot v_2) = 0 \mod 2 \,, \tag{4c}$$

$$\gcd(N_{i+2}, N_1)\ (W_i \cdot V_1) = 0 \mod 2 \,, \tag{4d}$$

$$\gcd(N_{i+2}, N_2)\ (W_i \cdot V_2) = 0 \mod 2 \,, \tag{4e}$$

$$N_{i+2} \left(W_i^2\right) = 0 \mod 2 \,, \tag{4f}$$

$$\gcd(N_{i+2}, N_{j+2}) \left(W_i \cdot W_j\right) = 0 \mod 2 \quad (i \neq j) \,, \tag{4g}$$

for $i, j = 1, \ldots, 6$ and where v_1 and v_2 denote the so-called twist vectors. They encode the geometrical rotation angles of a general $\mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2}$ orbifold geometry, while for $\mathbb{Z}_{N_1}$ orbifolds the twist $v_2 = (0^4)$ is not present. In addition, the gcd in eq. (4) denotes the greatest common divisor. These conditions are very restrictive and already forbid a huge fraction of points in the space $\mathbb{Z}^{128}$ corresponding to eq. (2). The only reasonable way to create a consistent gauge embedding matrix M is by successively creating shift vectors and Wilson lines step-by-step and checking each time if the relevant conditions from eqs. (4) are fulfilled for those combinations of shift vectors and Wilson lines that have been chosen so far, see Fig. 1.

In this paper we work out an extension of this logic, i.e. we create a successive search that only considers those areas in the heterotic orbifold landscape that can fulfill certain properties: first, in section 3 we will introduce phenomenological properties of the MSSM and then, in the main part of the paper in section 4, we define so-called contrast patterns that also can be checked at each step during the construction of a gauge embedding matrix M. By doing so, we will neglect those areas in the heterotic orbifold landscape that have no chance or an extremely low probability to host a valid MSSM-like orbifold model.

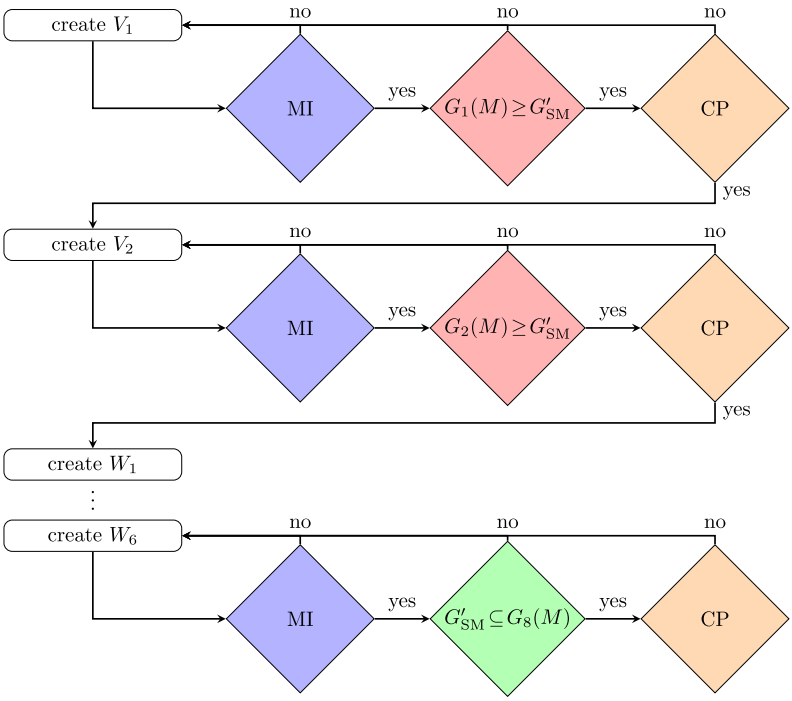


Fig. 1. Flowchart of the construction of shift vectors and Wilson lines, starting with the shift vector V_1 at step $n = 1$ and ending with the Wilson line W_6 at step $n = 8$. At each step $n = 1, \ldots, 8$, the vector M_n is chosen randomly and the corresponding modular invariance (MI) conditions are tested. If the vector passes this test, two additional conditions are applied: (i) As discussed in section 3.1 the gauge group $G_n(M)$ is computed using the already chosen vectors M_k for $k = 1, \ldots, n$. If $G_n(M)$ satisfies a necessary condition to host the non-Abelian gauge group factors $G'_{\rm SM} = {\rm SU}(3) \times {\rm SU}(2)$ of the SM in the first E_8 factor, i.e. $G_n(M) \geq G'_{\rm SM}$ in terms of their root systems, the model is passed on to the next condition. (ii) As introduced in section 4, the contrast patterns (CP) are imposed. Finally, after the last Wilson line $M_8 = W_6$ has been chosen successfully, the four-dimensional gauge group $G_{4D}(M) = G_8(M)$ must contain $G'_{\rm SM}$ in the first E_8 as described in section 3.2.

2.2. The Orbifolder

Given an orbifold geometry $\mathbb{O}$ and a consistent gauge embedding matrix M, it is in principle possible to compute the complete 4D orbifold model at low energies denoted by model(M).[2] In practice, some computations are too complicated, e.g. the strengths of Yukawa couplings. However, for a given M one can use the `orbifolder` [41][3] in order to get the massless string spectrum, denoted by spectrum(M), with all gauge charges. Moreover, the `orbifolder` can identify MSSM-like models, i.e. models with ${\rm SU}(3)_{\rm C} \times {\rm SU}(2)_{\rm L} \times {\rm U}(1)_Y$ gauge symmetry and the exact chiral spectrum of the MSSM plus at least one Higgs-pair and exotics that have to be vector-like with respect to the SM. Also discrete symmetries and a list of all allowed couplings up to a certain order in fields can be analyzed. In addition, the `orbifolder` can be used to

[2] If the orbifold geometry $\mathbb{O}$ is clear from the context, we will also name M as orbifold model.

[3] All 138 orbifold geometries with Abelian point groups and $\mathcal{N} = 1$ supersymmetry [37] are encoded in so-called geometry-files that can be read by the `orbifolder`. These geometry-files can be found as ancillary files to ref. [42].

identify inequivalent orbifold models by taking two orbifold models, model(M) and model(M'), to be equivalent if their massless spectra coincide (on the level of the non-Abelian representations and, in case of MSSM-like models, the $\mathrm{U}(1)_Y$ hypercharge), i.e.

$$\text{spectrum}(M) \;=\; \text{spectrum}(M') \quad \Rightarrow \quad \text{model}(M) \;\sim\; \text{model}(M')\,. \tag{5}$$

2.3. Searching in the Weyl chambers

A Weyl reflection of the gauge embedding vector M_k at the hyperplane orthogonal to the simple root α_I of $\mathrm{E}_8 \times \mathrm{E}_8$ is defined as

$$w_I(M_k) \;=\; M_k - (M_k \cdot \alpha_I)\,\alpha_I\,, \tag{6}$$

using $(\alpha_I)^2 = 2$ for $I = 1, \ldots, 16$. Then, it is easy to show that

$$w_I(M_k) \cdot w_I(M_\ell) \;=\; M_k \cdot M_\ell\,. \tag{7}$$

Hence, Weyl reflections leave the modular invariance conditions (4) invariant. Furthermore, one can show that they are symmetries of the full string theory

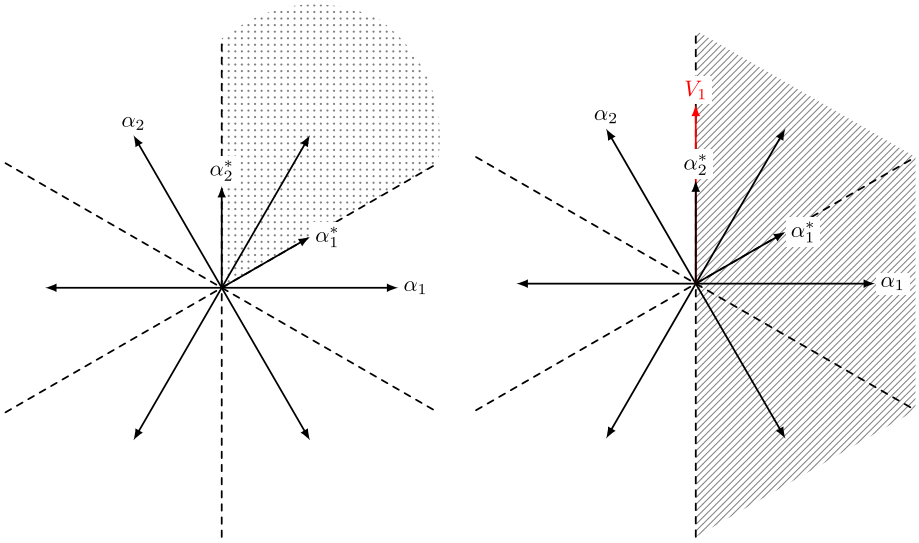
$$M' \;=\; w_I(M) \quad \Rightarrow \quad \text{model}\big(M\big) \;=\; \text{model}\big(M'\big)\,, \tag{8}$$

where w_I acts simultaneously on all shift vectors and Wilson lines encoded in M. Hence, the gauge embedding matrices M and $M' = w_I(M)$ are equivalent for all Weyl reflections. Now, Weyl reflections generate a group, the so-called Weyl group. For $\mathrm{E}_8 \times \mathrm{E}_8$, it has $\approx 5 \cdot 10^{17}$ elements. Consequently, the Weyl group of $\mathrm{E}_8 \times \mathrm{E}_8$ yields a huge redundancy between physically equivalent models in the heterotic orbifold landscape.

We can reduce the search space and therefore find more physically inequivalent models when we divide out this symmetry. For this task we propose a *fundamental Weyl chamber* search. The proposed technique is based on the algorithm of ref. [43] that any vector in the root space can be rotated to the fundamental Weyl chamber, which is defined as the subspace where all Dynkin labels $M_k \cdot \alpha_I$ are non-negative.

Starting from a gauge embedding matrix M we can imagine to apply the algorithm of ref. [43] such that the shift vector V_1 is rotated to the fundamental Weyl chamber, i.e. $V_1 \cdot \alpha_I \geq 0$ for all simple roots $I = 1, \ldots, 16$. Since we do not want to change the orbifold model by this transformation, we have to act with the same Weyl reflections that mapped V_1 to the fundamental Weyl chamber on the other vectors simultaneously. After this, we might still have some Weyl symmetries left, i.e. the shift vector V_1 may be invariant under certain Weyl reflections. These unbroken Weyl reflections are those that leave V_1 invariant, i.e. $w_I(V_1) = V_1$ if and only if $V_1 \cdot \alpha_I = 0$, i.e. $d_{1\,I} = 0$. These residual Weyl reflections can now be used to bring the next vector, in our case the Wilson line W_1, to an enlarged Weyl chamber which we define in analogy to the fundamental Weyl chamber but using only those Weyl reflections that leave V_1 invariant. Consequently, after the transformation of the Wilson line W_1 to the enlarged Weyl chamber, those Dynkin labels $W_1 \cdot \alpha_I$ are constrained to be non-negative that correspond to the Weyl reflections w_I that leave the shift vector V_1 invariant. This procedure can be reapplied to the next vectors until no Weyl symmetry is left.

The mindset above can be used to directly choose only gauge embedding matrices that solely have the first vector in the fundamental Weyl chamber and the following vectors are in the correspondingly enlarged versions of it, as illustrated in Fig. 2. To achieve this we expand our vectors not in the basis of the simple roots α_I but in the dual basis α_I^*, see eq. (2), where we can apply the



(a) Fundamental Weyl chamber of SU(3). (b) Domain for W_1 after a specific V_1 was chosen.

Fig. 2. These figures illustrate the algorithm to divide out the Weyl symmetry from the search process, exemplified at the root space of SU(3). In Fig. 2a the first vector V_1 can be restricted to lie in the fundamental Weyl chamber of SU(3) (shaded area) defined by the Weyl reflections w_1 and w_2. Thus, $d_{1\,I} \in \mathbb{N}_0$ for $I = 1, 2$. In Fig. 2b we have chosen a specific vector V_1 along the direction of α_2^* as an example. Consequently, the vector V_1 is invariant under the Weyl reflection w_1, i.e. $w_1(V_1) = V_1$, which corresponds to a vanishing Dynkin label $V_1 \cdot \alpha_1 = 0$. Hence, this choice for V_1 has not broken the whole Weyl symmetry and we can use the unbroken Weyl reflection w_1 to restrict the search space for W_1 to the enlarged Weyl chamber (shaded area) which is defined by $W_1 \cdot \alpha_1 \geq 0$. Hence, the coefficients of W_1 can be constrained as $d_{3\,1} \in \mathbb{N}_0$ and $d_{3\,2} \in \mathbb{Z}$. This procedure is continued for the next vectors and takes at each step all previously chosen vectors into account for computing the respective unbroken Weyl symmetry.

constraints on the Dynkin labels directly via the coefficients $d_{k\,I}$ of the gauge embedding matrix. For the first vector, i.e. the shift vector V_1, we have the full freedom of the Weyl group and can therefore choose this vector directly from the fundamental Weyl chamber

$$V_1 \cdot \alpha_I \;=\; \frac{d_{1\,I}}{N_1} \;\geq\; 0 \quad \Leftrightarrow \quad d_{1\,I} \;\in\; \mathbb{N}_0\,, \tag{9}$$

for $I = 1, \ldots, 16$. Thereafter, we have to compute the unbroken Weyl symmetry that can be exploited to restrict the second vector. This unbroken Weyl symmetry is generated by those Weyl reflections that leave V_1 invariant, i.e. V_1 has to be a fixed point of a Weyl reflection such that this Weyl reflection remains unbroken. Since we have chosen V_1 from the fundamental Weyl chamber it can only be a fixed point under a Weyl reflection if V_1 lies on the boundary of the fundamental Weyl chamber. This boundary is given by the union of the hyperplanes perpendicular to the simple roots, i.e. the hyperplanes at which the Weyl reflections w_I act [43]. Consequently, only those Weyl reflections w_I which leave all previously chosen vectors M_k invariant can still restrict the search space of the shift vector and Wilson lines that have to be chosen next. Therefore, at step n in Fig. 1 we can constrain the coefficients $d_{n\,I}$ of the vector M_n in eq. (2) as

$$d_{n\,I} \in \mathbb{N}_0 \quad \text{if } d_{k\,I} \;=\; 0 \quad \text{for all } k = 1, \ldots, n-1\,, \tag{10a}$$

$$d_{n\,I} \in \mathbb{Z} \quad \text{if } d_{k\,I} \;\neq\; 0 \quad \text{for any } k = 1, \ldots, n-1\,. \tag{10b}$$

Table 2
10^7 random models from the $\mathbb{Z}_6$-II orbifold geometry, ordered by their frequency of occurrence. We list only the first appearances of some special models according to the properties given in the last column. Note that the first column displays the ranking of the frequency of occurrence, where different types of models with the same frequency of occurrence appear at the same position in the ranking.

Ranking	Frequency of occurrence	Type of model
1	8008	gauge group $\mathrm{U}(1)^{16}$ (with 218 matter fields)
⋮	⋮	
19	1915	first non-Abelian gauge group (gauge group $\mathrm{SU}(2) \times \mathrm{U}(1)^{15}$)
⋮	⋮	
635	114	first gauge group $\mathrm{SU}(3) \times \mathrm{SU}(2)$
⋮	⋮	
739	10	first $\mathrm{SU}(3)_\mathrm{C} \times \mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_Y$ with 1 generation plus vector-like exotics
⋮	⋮	
747	2	first $\mathrm{SU}(3)_\mathrm{C} \times \mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_Y$ with 2 generations plus vector-like exotics
⋮	⋮	
748	1	first MSSM-like model with 3 generations plus vector-like exotics

3. Phenomenological constraints

Obviously, we have to neglect any orbifold model specified by a gauge embedding matrix M that does not obey the stringy consistency conditions on M: the geometrical constraints and the modular invariance conditions. Similarly, we can add phenomenologically inspired constraints on M: Any orbifold model whose four-dimensional gauge symmetry $G_{4D}(M)$ does not contain the one of the SM does not provide a valid model to describe nature. Importantly, if we search in the heterotic orbifold landscape taking only the stringy consistency condition into account, these phenomenologically uninteresting models build by far the main part of the heterotic orbifold landscape. We have verified this by constructing 10^7 random models in the $\mathbb{Z}_6$-II orbifold geometry that satisfy all stringy consistency conditions using the *fundamental Weyl chamber* search algorithm. These 10^7 models give rise to approximately $3.5 \cdot 10^6$ inequivalent massless spectra. Then, the inequivalent spectra are sorted according to their frequency of occurrence inside the full list of 10^7 random models. The result is shown in Fig. 3 and details on some of the inequivalent spectra are highlighted in Table 2. Consequently, from a phenomenological point of view the most uninteresting models turn out to have the highest repetition values, and the most interesting models are the rarest. As a remark, we cannot explain this imbalance, for example, by $\mathrm{E}_8 \times \mathrm{E}_8$ lattice translations and Weyl reflections. Since the models are compared on the level of their massless spectra, it is likely that a lot of these models actually differ by some additional model parameters, for instance by their Yukawa couplings. Nevertheless, we want to avoid these uninteresting models in our search for MSSM-like orbifold models. Therefore, we will describe in the upcoming sections how we can constrain our search to those areas of the heterotic orbifold landscape that can potentially host the SM.

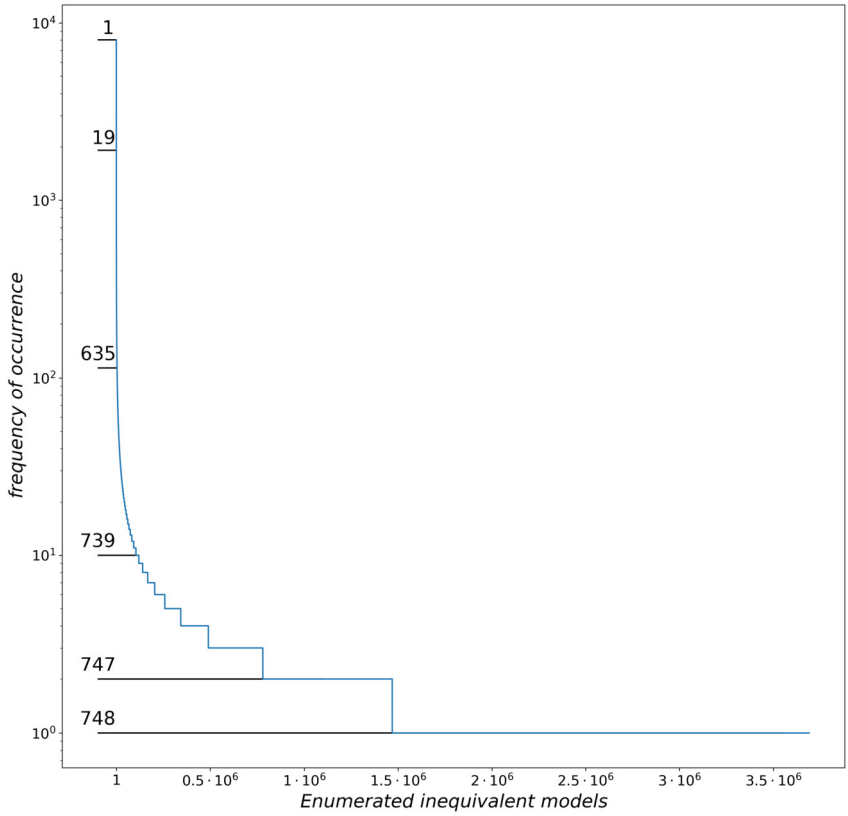


Fig. 3. Logarithmic plot of the frequency of occurrence of inequivalent $\mathbb{Z}_6$-II orbifold models. On the horizontal axis, the inequivalent models are enumerated from 1 to 3 690 513, while on the vertical axis we see the corresponding frequency of occurrence, i.e. model # 1 has a frequency of 8008, see also Table 2 for more details on some of these models. Moreover, the different frequencies of occurrence are ranked from position 1 to 748, where at rank 748 the first MSSM-like model appears with a frequency of occurrence of 1.

3.1. The pseudo-GUT constraint $G_n(M) \geq G'_{\rm SM}$

We want to avoid to produce phenomenologically uninteresting models that have a gauge group smaller than the SM gauge group factors $G'_{\rm SM} = {\rm SU}(3) \times {\rm SU}(2)$.[4] Hence, we can check how far the gauge group $G_n(M)$ is already broken at each step $n = 1, \ldots, 8$ in the algorithm illustrated in Fig. 1. Since additional shift vectors or Wilson lines can only break the gauge group further, i.e. $G_{n+1}(M) \subseteq G_n(M)$, the SM gauge group provides us with a lower bound on the breaking pattern at each step n. A fast way to check the size of the remaining gauge group $G_n(M)$ is to compute the number of unbroken roots $N_{\rm ur}$ after the first n vectors M_k, for $k = 1, \ldots, n$, have been chosen to specify a consistent gauge embedding matrix M,

$$N^{(\alpha)}_{\rm ur}(M) \;=\; \sum_{p \,\in\, \Phi_{\rm E_8}} \left(\prod_{k\,=\,1}^{n} \delta\left(p \cdot M^{(\alpha)}_k \right) \right) , \tag{11}$$

[4] Note that the existence of an anomaly-free ${\rm U}(1)_Y$ hypercharge can and will be tested only at the very end of the search algorithm by the `orbifolder`, after the full orbifold model has been specified.

for $\alpha = 1, 2$ corresponding to both E_8 factors and Φ_{E_8} is defined as the root system of E_8 with 240 roots p. Furthermore, the subindex "ur" in eq. (11) denotes *unbroken roots* and $\delta(x) = 1, 0$ depending on whether x is integer or not, respectively. In the case of the SM we have six unbroken roots for SU(3) and two unbroken roots for the SU(2) factor, i.e. $N_{\rm ur}(\text{SM}) = 8$. This is our lower bound at each step n in the production of our model M for the first E_8 factor, i.e.

$$N_{\rm ur}^{(1)}(M) \;\geq\; 8 \quad \text{at each step } n \,. \tag{12}$$

On the other hand, the second E_8 factor is free to produce any additional hidden gauge groups. Note that $\text{SU}(2) \times \text{SU}(2) \times \text{SU}(2) \times \text{SU}(2)$ could fulfill the constraint (12) but is already broken too far. Due to this we demand in addition that one gauge factor of $G_n(M)$ has a root system that allows for SU(3), i.e. with six or more unbroken roots.[5] If a newly chosen vector M_k results in a gauge group breaking below these lower bounds, we neglect this vector and choose the same vector again, until it fulfills this constraint. In the following we call this constraint the *pseudo-GUT* constraint.

3.2. *The Standard Model gauge group constraint:* $\text{SU}(3) \times \text{SU}(2) \subseteq G_{4D}(M)$

The *pseudo-GUT* constraint is a necessary condition for a model to contain the non-Abelian gauge group factors $G'_{\rm SM} = \text{SU}(3) \times \text{SU}(2)$ of the SM at each step of the construction of the gauge embedding matrix M. However, our search focuses on MSSM-like models with $\text{SU}(3)_{\rm C} \times \text{SU}(2)_{\rm L} \times \text{U}(1)_Y$ gauge symmetry in 4D and not on grand unified models like Pati-Salam or SU(5). Hence, after we have chosen the last vector M_k of our orbifold model (taking the geometrical constraints into account), we can check that the model M has a 4D gauge symmetry

$$G_{4D}(M) \;=\; \text{SU}(3) \times \text{SU}(2) \times G_{\rm hidden} \,. \tag{13}$$

We denote this constraint by $\text{SU}(3) \times \text{SU}(2) \subseteq G_{4D}(M)$.

Due to the geometrical constraints, we first have to identify the last shift vector or Wilson line that can be chosen independently, i.e. which is not of order one and not equal to a previous vector. For example, for the $\mathbb{Z}_6$-II orbifold geometry this results in the Wilson line W_6, as can be seen in Table 2. However, note that for some other orbifold geometries like $\mathbb{Z}_3 \times \mathbb{Z}_6$ (2, 2) all Wilson lines are fixed by the geometry, $W_i = (0^{16})$ for $i = 1, \ldots, 6$, and the second shift vector V_2 has to enable the $G_{4D}(M)$ constraint. The constraint is checked by computing the unbroken roots from the first E_8 factor and the sizes of their orthogonal root systems. This means that in order to contain $\text{SU}(3) \times \text{SU}(2)$ at least two root systems, one of size six and another of size two, have to be present, where we allow for additional gauge group factors, also from the first E_8 factor.

We implement the phenomenological constraints from section 3.1 and section 3.2 into our search algorithm and apply it to the test case of $\mathbb{Z}_6$-II orbifold models. It turns out that the probability of finding an MSSM-like model increases by a factor 10 from $\frac{1}{10\,000\,000} = 10^{-7}$ in the case without the phenomenological constraints to $\frac{3}{2\,665\,463} \approx 10^{-6}$ in the case where the phenomenological constraints are applied.[6] In addition, we use physical intuition that MSSM-like

[5] There exist also the special cases SO(8) and SU(4), which can not be broken to $G'_{\rm SM}$. However, they occur on rare occasions and in order to keep our search fast, we avoid a separate check. Moreover, these cases are excluded from the search by the upcoming constraint in section 3.2.

[6] We compute the probability on basis of the equivalent models in order to avoid "floating correlations" [44].

Table 3
Datasets created in the $\mathbb{Z}_6$-II orbifold landscape. Each dataset incorporates all of the constraints, indicated by the names in the first column, of the datasets above. The *fundamental Weyl chamber* dataset utilizes the Weyl symmetry, see section 2.3, while the *phenomenology* dataset makes additionally use of the constraints developed in section 3. In the next rows, we apply our contrast patterns: first, we demand $N_{\text{ur}}^{(2)} \geq 6$ from section 4.2.1 and obtain the *hidden* E_8 dataset. Note that in the dynamic search we modify this constraint as explained in section 4.2.2 to $N_{\text{ur}}^{(2)} \geq X$ for $X \in \{8, 10, 12, \ldots, 86\}$ and obtain the *dynamic hidden* E_8 dataset. Here, the case $X = 6$ was disregarded since it was already sampled in the *hidden* E_8 dataset. Finally, the *U-sector* dataset was created using the *U-sector* contrast pattern from section 4.2.3 in addition. Note that we also made use of the additional conditions $W_5 = (0^{16})$ or $W_3 = W_4 = (0^{16})$, where $W_5 = (0^{16})$ is known to be beneficial for finding MSSM-like models, see ref. [20].

	Dataset	Condition	# models	# MSSM-like	# inequiv. MSSM-like		
traditional	*fundamental Weyl chamber*		10 000 000	1	1		
	phenomenology		2 665 463	3	3	130	
		$W_5 = (0^{16})$	2 551 272	509	129		
contrast patterns	*hidden* E_8		2 543 415	12	11	136	415
		$W_5 = (0^{16})$	2 609 872	863	135		
	dynamic hidden E_8		1 876 273	3299	245	395	468
		$W_5 = (0^{16})$	1 231 608	8455	321		
		$W_3 = (0^{16})$	378 604	7	2		
	U-sector		4 793 146	4953	357	459	
		$W_5 = (0^{16})$	3 046 262	17 406	358		

models are often related to a vanishing Wilson line [20] and perform a second search where we set $W_5 = (0^{16})$ by hand. The results are summarized in Table 3 (where the corresponding dataset is called *phenomenology*).

4. Contrast patterns for $\mathbb{Z}_6$-II orbifolds

In the previous section, we discussed phenomenological constraints that can be checked easily during the search for MSSM-like orbifold models. Importantly, these conditions are absolutely necessary for a model to be MSSM-like (but not sufficient). Now, we want to extend this procedure to include additional constraints (so-called contrast patterns) for MSSM-like models by exploiting techniques from data mining. These new constraints will be determined by a statistical approach. Hence, demanding them can potentially rule out a few MSSM-like models. In other words, the new constraints are not necessarily satisfied for all MSSM-like models but they significantly enhance the probability for a given model to be MSSM-like. In this way, we will constrain the heterotic orbifold landscape further to the areas of MSSM-like models. Some of these areas are hardly accessible by the conventional search algorithm but easy to access due to the significantly enhanced probability given the additional constraints from the contrast patterns.

A contrast pattern c can be defined as a pattern whose supports differ significantly among the datasets under contrast [45]. Here, the support is defined as

$$\text{supp}(c, D) = \frac{|\{M \in D \mid M \text{ satisfies } c\}|}{|D|}, \tag{14}$$

where D is a set of data points, i.e. orbifold models, and c is a set of certain constraints that have to be fulfilled. In our case, we have two datasets that are under contrast: $D_{\text{MSSM-like}}$ and

$D_{\not{\text{MSSM}}\text{-like}}$, which is the set of MSSM-like models and the complementary set, respectively. In other words, we are searching for constraints c that are satisfied for nearly all MSSM-like models while they are violated by a huge fraction of $\not{\text{MSSM}}$-like models. In the ideal case, we can identify contrast patterns c with $\text{supp}(c, D_{\text{MSSM-like}}) = 1$ and $\text{supp}(c, D_{\not{\text{MSSM}}\text{-like}}) = 0$. This can be formalized by defining the growth rate

$$\text{gr}(c, D_{\text{MSSM-like}}, D_{\not{\text{MSSM}}\text{-like}}) \;=\; \frac{\text{supp}(c, D_{\text{MSSM-like}})}{\text{supp}(c, D_{\not{\text{MSSM}}\text{-like}})}\,, \tag{15}$$

which has to be maximized. In the following, we will often just write $\text{gr}(c)$ if the datasets $D_{\text{MSSM-like}}$ and $D_{\not{\text{MSSM}}\text{-like}}$ are clear from the context.

To understand the growth rate better and get some intuition for its value, we rewrite it in terms of the probability $\hat{p}$. Here, the hat indicates that we estimate the probability by the sample proportion $\hat{p}(Y) = {}^{N_Y}/_{N}$, where $Y \in \{\text{MSSM-like}, \not{\text{MSSM}}\text{-like}\}$ and the total sample size is given by $N = N_{\text{MSSM-like}} + N_{\not{\text{MSSM}}\text{-like}}$. Then, one finds

$$\text{gr(c)} = \frac{\hat{p}^c(\text{MSSM-like})}{\hat{p}(\text{MSSM-like})}\,\frac{\hat{p}(\not{\text{MSSM}}\text{-like})}{\hat{p}^c(\not{\text{MSSM}}\text{-like})}\,, \tag{16}$$

where $\hat{p}^c(Y) = {}^{N^c_Y}/_{N^c}$ with $N^c_Y = |\{M \in D_Y \mid M \;\text{ satisfies }\; c\}|$ is the probability of a model being $Y = \text{MSSM-like}$ or $Y = \not{\text{MSSM}}\text{-like}$ given the constraints c and $\hat{p}(Y)$ is the corresponding probability without imposing the constraints c. Then, one can solve eq. (16) for $\hat{p}^c(\text{MSSM-like})$ as a function of $\hat{p}(\text{MSSM-like})$ as follows

$$\hat{p}^c(\text{MSSM-like}) \;=\; \frac{\text{gr(c)}\,\hat{p}(\text{MSSM-like})}{1 + (\text{gr(c)} - 1)\,\hat{p}(\text{MSSM-like})}\,. \tag{17}$$

Here, one can observe several cases:

$$\hat{p}(\text{MSSM-like}) \ll 1 \;:\; \hat{p}^c(\text{MSSM-like}) \;=\; \text{gr(c)}\hat{p}(\text{MSSM-like}) + \mathcal{O}\Big(\hat{p}(\text{MSSM-like})^2\Big)\,, \tag{18a}$$

$$\text{gr(c)} \;=\; 1 \;:\; \hat{p}^c(\text{MSSM-like}) \;=\; \hat{p}(\text{MSSM-like})\,, \tag{18b}$$

$$\text{gr(c)} \;=\; 0 \;:\; \hat{p}^c(\text{MSSM-like}) \;=\; 0\,, \tag{18c}$$

$$\text{gr(c)} \;\to\; \infty \;:\; \hat{p}^c(\text{MSSM-like}) \;\to\; 1\,, \tag{18d}$$

where the Taylor expansion in eq. (18a) converges for $\hat{p}(\text{MSSM-like}) < \frac{1}{|\text{gr}(c)-1|}$. Now, eq. (18a) can be interpreted easily: For $\text{gr}(c) < 1$ we have a negative effect on our favored class of MSSM-like models. For $\text{gr}(c) = 1$ the effects on both classes cancel each other and for $\text{gr}(c) > 1$ we have a positive effect, i.e. a higher probability to find MSSM-like models in the subspace defined by the contrast patterns c.

However, before we can start to search for contrast patterns c, we have to define some (physical) quantities that possibly can lead to such patterns. This is known as feature engineering as explained in the next section.

4.1. Feature engineering

In this section, we will define (physical) quantities for a given orbifold model M. In the context of data mining, we will call such quantities *features* and their construction is called *feature engineering*. In general terms, feature engineering denotes the process of computing useful quantities from the raw data. For example, neural networks generate features in each hidden layer on

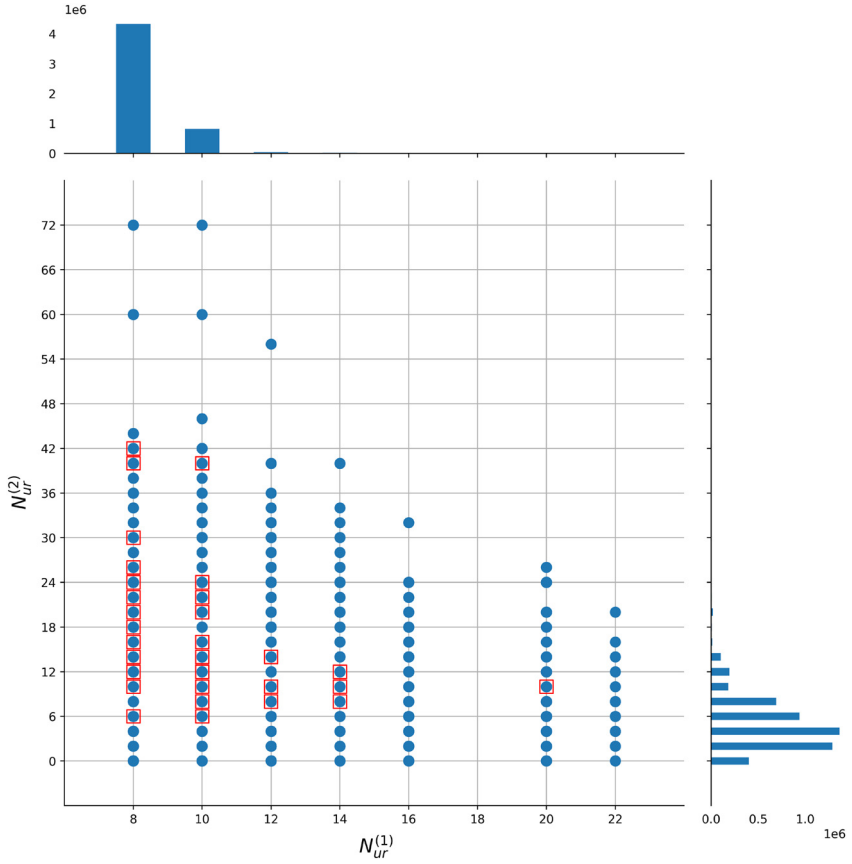


Fig. 4. Jointplot (see ref. [46]) of $N_{\text{ur}}^{(1)}$ and $N_{\text{ur}}^{(2)}$ for the complete *phenomenology* dataset of $\mathbb{Z}_6$-II orbifold models, see Table 3. Due to the constraint $\text{SU}(3) \times \text{SU}(2) \subseteq G_{4D}(M)$ from section 3.2 we have a bound $N_{\text{ur}}^{(1)} \geq 8$ in this dataset. The central part shows a scatter plot, where ~~MSSM~~-like and MSSM-like orbifold models are marked by blue circles and red boxes, respectively. Above and on the right hand side of the scatter plot we give the histogram on the number of models for given values of $N_{\text{ur}}^{(1)}$ and $N_{\text{ur}}^{(2)}$, respectively. Note that the number of MSSM-like models is negligible compared to the ~~MSSM~~-like models in this dataset. Hence, the histograms visualize the frequency of occurrence for ~~MSSM~~-like models only. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

their own during training. However, this is one of the biggest open problems of neural networks: it is in general very difficult to extract any meaning of the features that a neural network has learned on its own – these features hardly yield any knowledge gain. Alternatively, in this paper we will use physical intuition and knowledge of the system to create features. Then, after visualizing some of these features, we can use machine learning techniques (i.e. a decision tree) to quantify if our educated guess for a certain feature, or combinations of multiple features, leads to a correlation between these features and the property of M being MSSM-like or not. If such a correlation exists (i.e. if $\text{gr}(c) > 1$), we have identified a promising contrast pattern c. Moreover, if we can check this pattern c easily in the search algorithm displayed in Fig. 1, we can utilize it to reduce the search space in the heterotic orbifold landscape to areas where MSSM-like orbifold models accumulate. The advantage of this approach is obvious: by construction we have a straightforward physical interpretation of our features.

However, in general it is very difficult to identify useful features. In our case, we have tried many concepts, for instance, local GUTs, the breaking patterns of each shift vector and each Wilson line, the breaking patterns of certain combinations thereof, the number of non-Abelian gauge group factors in 4D, and many more. We know from section 2 that the 4D gauge group has a great impact and it can be checked easily at every step of the production of a model. Therefore, there is hope that additional features can be found from the 4D gauge group. When attempting to identify a promising feature, data visualization can be very useful: In Fig. 4, we plot the number of unbroken roots $N_{\text{ur}}^{(\alpha)}$ from each E_8 factor in a scatter plot against each other using the *phenomenology* dataset created in section 3. In addition, the respective histograms for the number of orbifold models with a certain number of unbroken roots are displayed in Fig. 4 for each axis (i.e. for each E_8 factor). To see if the feature $N_{\text{ur}}^{(\alpha)}$ has the potential to be useful as a contrast pattern, we have to pay attention to two aspects of this plot. First, we have to identify areas in this plot where no MSSM-like orbifold model is present. Second, it is important that such an area is not only qualitatively separated from MSSM-like orbifold models but also quantitatively interesting, i.e. highly populated with ~~MSSM~~-like orbifold models. This aspect can be read off from the respective histogram in Fig. 4. By doing so, we identify the area $N_{\text{ur}}^{(2)} < 6$ that does not contain any MSSM-like $\mathbb{Z}_6$-II orbifold model but is fairly high populated with ~~MSSM~~-like models in our *phenomenology* dataset. Hence, the condition $N_{\text{ur}}^{(2)} < 6$ is our first promising candidate for a contrast pattern and we expect to get a strong reduction of the $\mathbb{Z}_6$-II orbifold landscape by excluding this area from the search. In contrast, an example of an area that consists of ~~MSSM~~-like $\mathbb{Z}_6$-II models only but is irrelevant due to the small number of models is given by $N_{\text{ur}}^{(2)} > 42$ or $N_{\text{ur}}^{(1)} > 20$: the respective histograms (in Fig. 4, the vertical histogram on the right-hand side for $N_{\text{ur}}^{(2)} > 42$ and the horizontal histogram above the scatter plot for $N_{\text{ur}}^{(1)} > 20$) show that the number of $\mathbb{Z}_6$-II orbifold models in these regions is very small.

In addition to the above features, we will also use the numbers of orbifold-invariant bulk matter fields as additional features. They are computed similar to the number of unbroken roots of the gauge group in eq. (11), with the difference of an additional displacement from the geometrical twist vector v, i.e. at each step n of our search algorithm displayed in Fig. 1 we compute

$$N_{\text{U}_a}^{(\alpha)}(M) \;=\; \sum_{p \,\in\, \Phi_{\text{E}_8}} \prod_{k\,=\,1}^{n} \delta\left(p \cdot M_k^{(\alpha)} - \Theta(2-k)\, q_{(a)} \cdot v_{(k)} \right) , \tag{19}$$

for $\alpha = 1, 2$ and $a = 1, 2, 3$. Note that the term $q_{(a)} \cdot v_{(k)}$ in eq. (19) is turned off for the Wilson lines $M_k^{(\alpha)}$, $k = 3, \ldots, 8$, using

$$\Theta(x) \;=\; \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 & \text{if} \quad x \geq 0\,. \end{cases} \tag{20}$$

Furthermore, the vectors $q_{(1)} = (0, -1, 0, 0)$, $q_{(2)} = (0, 0, -1, 0)$ and $q_{(3)} = (0, 0, 0, -1)$ give rise to the three untwisted sectors U_a, $a = 1, 2, 3$, and correspond to the three directions of the internal vector-boson index of the ten-dimensional $\text{E}_8 \times \text{E}_8$ gauge bosons, respectively. Finally, the twist vectors in eq. (19) are given by $v_{(1)} = (0, \frac{1}{6}, \frac{1}{3}, -\frac{1}{2})$ and $v_{(2)} = (0^4)$ is not present for the $\mathbb{Z}_6$-II orbifold geometry. Due to the δ-condition in eq. (19) all features $N_{\text{U}_a}^{(\alpha)}(M)$ and $N_{\text{ur}}^{(\alpha)}(M)$ are independent.

Table 4
Overview table of the features that we use for contrast mining. Each feature is evaluated for all orbifold models M of the dataset under investigation.

	First E_8				Hidden E_8			
feature	$N_{\text{ur}}^{(1)}$	$N_{\text{U}_1}^{(1)}$	$N_{\text{U}_2}^{(1)}$	$N_{\text{U}_3}^{(1)}$	$N_{\text{ur}}^{(2)}$	$N_{\text{U}_1}^{(2)}$	$N_{\text{U}_2}^{(2)}$	$N_{\text{U}_3}^{(2)}$

4.2. *Decision tree and* false negatives

In this section, we will use a decision tree [47] in order to identify those features from Table 4 that correlate with the property of a model being MSSM-like or not. If such a correlation exists, the corresponding feature can be used as contrast pattern in our search algorithm for MSSM-like orbifold models. A decision tree belongs to the class of supervised machine learning and can be used for the purpose of classification or regression. In our case, we want to classify whether a given orbifold model is MSSM-like or not using simple true-or-false decisions on the features listed in Table 4. Then, by analyzing the decisions made inside of the decision tree, we can identify those features that lead to successful contrast patterns. This is possible since a decision tree partitions the feature space along linear decision boundaries, which are orthogonal to the feature axes. Even though this limits the power of classification, it has the advantage that the resulting contrast patterns can be interpreted easily.

In more detail, our decision tree is a function from some of the features listed in Table 4 to the classification value denoted by Y, i.e. it is of the form

$$\left(N_{\text{ur}}^{(1)}(M), N_{\text{U}_1}^{(1)}(M), N_{\text{U}_2}^{(1)}(M), N_{\text{U}_3}^{(1)}(M), \dots, \right) \mapsto Y_{\text{predicted}}(M) \,, \tag{21}$$

where $Y_{\text{predicted}}(M) \in \{\text{MSSM-like}, \cancel{\text{MSSM}}\text{-like}\}$, and it can be applied to all orbifold models M. Note that we can compute for each orbifold model M both, the features and the correct classification value $Y_{\text{correct}}(M) \in \{\text{MSSM-like}, \cancel{\text{MSSM}}\text{-like}\}$ using the `orbifolder`. Yet, the benefit of using a decision tree is given by the possibility to uncover unknown correlations between our features and the property $Y_{\text{correct}}(M)$ of an orbifold model M to be MSSM-like or not. Furthermore, it is by no means guaranteed that a function like eq. (21) exists. We will only know about its existence after we have trained and tested our decision tree.

In a first step, the decision tree has to be trained, i.e. the algorithm tries to learn the function (21). To do so, it needs a training set, i.e. a list of orbifold models, where for each orbifold model M_i we have computed the values of our features and the correct classification values $Y_{\text{correct}}(M_i)$ using the `orbifolder`. More explicitly, the training set reads

$$\begin{aligned}\text{training set} = \{\{&N_{\text{ur}}^{(1)}(M_1), N_{\text{U}_1}^{(1)}(M_1), N_{\text{U}_2}^{(1)}(M_1), N_{\text{U}_3}^{(1)}(M_1), \dots, Y_{\text{correct}}(M_1)\}, \\ &\{N_{\text{ur}}^{(1)}(M_2), N_{\text{U}_1}^{(1)}(M_2), N_{\text{U}_2}^{(1)}(M_2), N_{\text{U}_3}^{(1)}(M_2), \dots, Y_{\text{correct}}(M_2)\}, \dots\} \,,\end{aligned} \tag{22}$$

and during training the decision tree tries to adjust the function eq. (21) such that $Y_{\text{predicted}}(M) = Y_{\text{correct}}(M)$ for all models M from the training set. After training, we can evaluate the trained decision tree (21) on the so-called validation set

$$\begin{aligned}\text{validation set} = \{\{&N_{\text{ur}}^{(1)}(P_1), N_{\text{U}_1}^{(1)}(P_1), N_{\text{U}_2}^{(1)}(P_1), N_{\text{U}_3}^{(1)}(P_1), \dots, Y_{\text{correct}}(P_1))\}, \\ &\{N_{\text{ur}}^{(1)}(P_2), N_{\text{U}_1}^{(1)}(P_2), N_{\text{U}_2}^{(1)}(P_2), N_{\text{U}_3}^{(1)}(P_2), \dots, Y_{\text{correct}}(P_2)\}, \dots\} \,,\end{aligned} \tag{23}$$

containing the features of some other orbifold models P_i. Then, we can compare the results $Y_{\text{predicted}}(P_i)$ of the decision tree (21) to the correct values $Y_{\text{correct}}(P_i)$ obtained from the `orbifolder`. In this way, we can check whether our decision tree was able to identify the function eq. (21) between our features and the property of a model being MSSM-like or not.

In practice, a decision tree will not be trained perfectly. First of all, it is possible that there is no exact functional dependency of the form eq. (21). Furthermore, even in the case when such a functional dependency would exist in principle, the decision tree might be unable to learn it, possibly because the training set was too small or imbalanced. In general, we can distinguish between two types of errors, i.e. cases where $Y_{\text{correct}}(M) \neq Y_{\text{predicted}}(M)$. They are called:

- *false positives*: $Y_{\text{correct}}(M) = \not{\text{MSSM}}$-like but $Y_{\text{predicted}}(M) =$ MSSM-like
- *false negatives*: $Y_{\text{correct}}(M) =$ MSSM-like but $Y_{\text{predicted}}(M) = \not{\text{MSSM}}$-like

Every classification process tries to minimize the number of false predictions. However, at a certain level it always comes to a trade-off between *false positives* and *false negatives* and we have to decide whether we want to suppress one of them for the drawback of raising the other one. In our case, a *false positive* classification by the decision tree is not a big problem since we can simply check each of these orbifold models afterwards explicitly using the `orbifolder`. However, in the case of a *false negative* classification the consequences are that we will loose an MSSM-like orbifold model. Since MSSM-like orbifold models are far too valuable to us, we want to minimize the number of *false negative* cases by all means, while we want to keep the number of *false positives* as low as possible. Therefore, we introduce a loss matrix L, which informs the machine learning algorithm about the different importance of certain models [48]. We choose a loss matrix

$$L = \begin{pmatrix} 0 & 10^6 \\ 1 & 0 \end{pmatrix} , \tag{24}$$

where the two rows correspond to the correct value $Y_{\text{correct}}(M)$ being either MSSM-like or not, and the two columns correspond to the predicted values $Y_{\text{predicted}}(M)$ being either MSSM-like or not. Then, L_{12} corresponds to the *false negative* cases of an MSSM-like orbifold model M that has been classified by the decision tree to be $Y_{\text{predicted}}(M) = \not{\text{MSSM}}$-like. As this is very undesirable, the system is punished with a large loss value $L_{12} = 10^6$. As discussed before, the other possible error of a *false positive* classification is not so severe. Hence, we set $L_{21} = 1$. This will guide the decision tree algorithm towards suppressing the *false negative* cases such that we do not miss any MSSM-like orbifold models.

For later convenience, we will quantify the quality of the predictions by the recall. It is defined as the number of correct predictions of the MSSM-like class divided by the total number of MSSM-like orbifold models. Hence, if the number of *false negatives* for all MSSM-like orbifold models P_i is zero, the recall is 1.00 on the validation set and all MSSM-like orbifold models are assigned with the correct value $Y =$ MSSM-like.

In the following, we apply decision trees to our features in order to extract promising contrast patters.

4.2.1. The hidden E_8 *contrast pattern*

As a first step, we have to define our datasets. The training and validation set are created by a random split (using a validation size of 33%) of the *phenomenology* dataset from Table 3 (based on the phenomenological constraints from section 3.1 and section 3.2). However, we add a small

modification to the dataset to avoid *data leakage*. Data leakage refers to the mistake to inform the machine learning algorithm about data from the validation set during training. In this case, the machine learning algorithm might overfit on some of the data from the validation set even though the data was divided into training and validation set. As the performance of a machine learning model on the validation set is a measure for its ability to generalize to unseen data, this mistake has to be avoided. In our case this could happen if, for example, there exists an MSSM-like model that completely dominates all MSSM-like models with all of its equivalent copies. Then, this model would appear most likely in both, training and validation set. Hence, the machine learning algorithm would see this model during training. Moreover, the same model would dominate the results on the validation set and pretend that the learned predictions generalize to generic MSSM-like models. Therefore, to avoid data leakage we only use inequivalent MSSM-like models. Nevertheless, for the $\not{\text{MSSM}}$-like models we keep the equivalent models, since the frequency of occurrence gives us a notion of the size of the area that a certain split in the decision tree excludes. In more detail, we have to perform our search for MSSM-like orbifold models in the space $\mathbb{Z}^{128}$ of gauge embedding matrices $\{M_i\}$, see eq. (2), even though we are interested in the space of physically inequivalent models $\{\text{model}(M_j)\}$, or more precisely, in the space of inequivalent massless particle spectra $\{\text{spectrum}(M_k)\}$. Now, we defined features that directly depend on $\text{spectrum}(M)$, not on $d \in \mathbb{Z}^{128}$ and our decision tree performs its splits based on these features. Consequently, a certain split in the decision tree will exclude all points $d \in \mathbb{Z}^{128}$ that give rise to the same excluded features. In this way, a small restriction in feature space gives rise to a huge effect in the space $\mathbb{Z}^{128}$ of gauge embedding matrices. Moreover, by not restricting the feature space too much, we leave enough room to discover new MSSM-like models, also in unexpected areas of the landscape. In contrast, if we had used only inequivalent $\not{\text{MSSM}}$-like models for the training of our decision tree, the decision tree would not care to exclude a single model even though it might actually correspond to a huge area in $\mathbb{Z}^{128}$. At the end, we want to enhance the search algorithm. Hence, it is better to exclude a few models with extraordinary high frequency of occurrence than multiple models with very low one.

Now, we train the decision tree on the training set. Here, we tune the hyperparameters such that we get a recall value for MSSM-like models of 1.00 on the validation set. This is due to the fact that we want to find contrast patterns that are satisfied by all MSSM-like models contained in the validation set, e.g. lowering the loss value L_{12} to 10^5 already leads to undesirable *false negatives*. During training, the decision tree identifies areas in feature space and assigns the two classes {MSSM-like, $\not{\text{MSSM}}$-like} to them, using the data from the training set. However, we want the decision tree to assign the class $\not{\text{MSSM}}$-like only to those areas that are also highly populated with $\not{\text{MSSM}}$-like models. In this way, we can be sure that the probability for an MSSM-like model is extremely small in these areas of $\not{\text{MSSM}}$-like models. This can be achieved using a technique called pruning. Consequently, the complexity of the decision tree is reduced to a minimum and we keep the possibility to find MSSM-like models in rather unexpected areas of the landscape.

The resulting decision tree is displayed in Fig. 5. We find that the pattern obtained by intuition from the scatter plot Fig. 4 is actually the strongest pattern in the full feature space. This pattern is given by a lower bound on the number of unbroken roots from the hidden E_8 factor: From the second node in the second line of the decision tree we can extract the condition

$$c_{\text{hidden E}_8} = \left\{ N_{\text{ur}}^{(2)}(M) \geq 6 \right\}, \tag{25}$$

for a model M to have a high probability to be MSSM-like. We call this condition the *hidden* E_8 contrast pattern. In principle, our decision tree contains additional splits. However, we want to stay conservative with our search and avoid too enthusiastic splits. Hence, we stay with the first

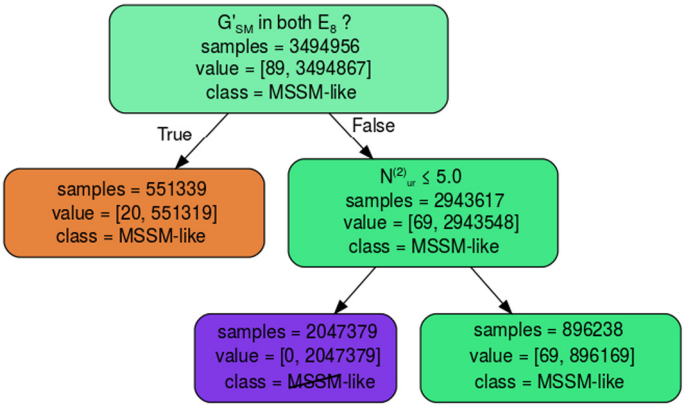


Fig. 5. Decision tree on the *phenomenology* dataset from Table 3, evaluated on the training set. We can extract the contrast pattern $N_{\text{ur}}^{(2)} \leq 5$ for ~~MSSM~~-like models from this tree, which we reformulate into a positive condition $N_{\text{ur}}^{(2)} \geq 6$ for MSSM-like models. Let us explain the details at the example of the uppermost node. This node contains: (i) the condition G'_{SM} *in both* E_8*?* that has to be evaluated, (ii) *samples=3 494 956* gives the total number of models in this node, (iii) *value* = [89, 3 494 867] gives the number of MSSM-like and ~~MSSM~~-like models in this node, respectively, and, finally, (iv) *class* = *MSSM-like* is the prediction for all models in this intermediate node. The final prediction for the models is given by the leaf nodes.

and most important split for now. Based on the training set from our *phenomenology* dataset we can estimate the growth rate of the *hidden* E_8 contrast pattern to be

$$\text{gr}\left(c_{\text{hidden E}_8}, D_{\text{MSSM-like}}, D_{\text{~~MSSM~~-like}}\right) = \frac{1}{\frac{896\,169+551\,319}{3\,494\,867}} \approx 2.4 > 1\,, \tag{26}$$

using eq. (15) with $\text{supp}(c_{\text{hidden E}_8}, D_{\text{MSSM-like}}) = 1$ for our *phenomenology* dataset and the numbers for $\text{supp}(c_{\text{hidden E}_8}, D_{\text{~~MSSM~~-like}})$ can be read off from Fig. 5. In other words, we can modify our search algorithm such that we would have avoided 2 047 379 ~~MSSM~~-like models in the training set. Hence, the contrast pattern (25) allows us to exclude a huge area in the $\mathbb{Z}_6$-II orbifold landscape which statistically does not lead to MSSM-like models. Instead, we can invest the gained computing time to search in areas where the probability of a model to be MSSM-like is significantly increased.

We implement the *hidden* E_8 contrast pattern into our search algorithm displayed in Fig. 1 and perform an intensive search using in addition different constraints on the Wilson lines: First, we allow all Wilson lines to be non-trivial and then, motivated by ref. [20], we turn off W_5 by hand. The resulting dataset is called *hidden* E_8 and summarized in Table 3. One observes an increase of the probability to find an MSSM-like model from

$$\hat{p}^{\text{phenomenology}}(\text{MSSM-like}) = \frac{512}{5\,216\,735} \approx 10^{-4} \quad \text{to} \tag{27a}$$

$$\hat{p}^{\text{hidden E}_8}(\text{MSSM-like}) = \frac{875}{5\,153\,287} \approx 2 \cdot 10^{-4}\,, \tag{27b}$$

which is consistent with the estimated growth rate in eq. (26). However, these are the probabilities to find any MSSM-like model and it does not need to be inequivalent to the already known ones. Unfortunately, the total number of inequivalent MSSM-like models did not satisfy our expectations: it increased from 130 inequivalent MSSM-like models in the *phenomenology* dataset

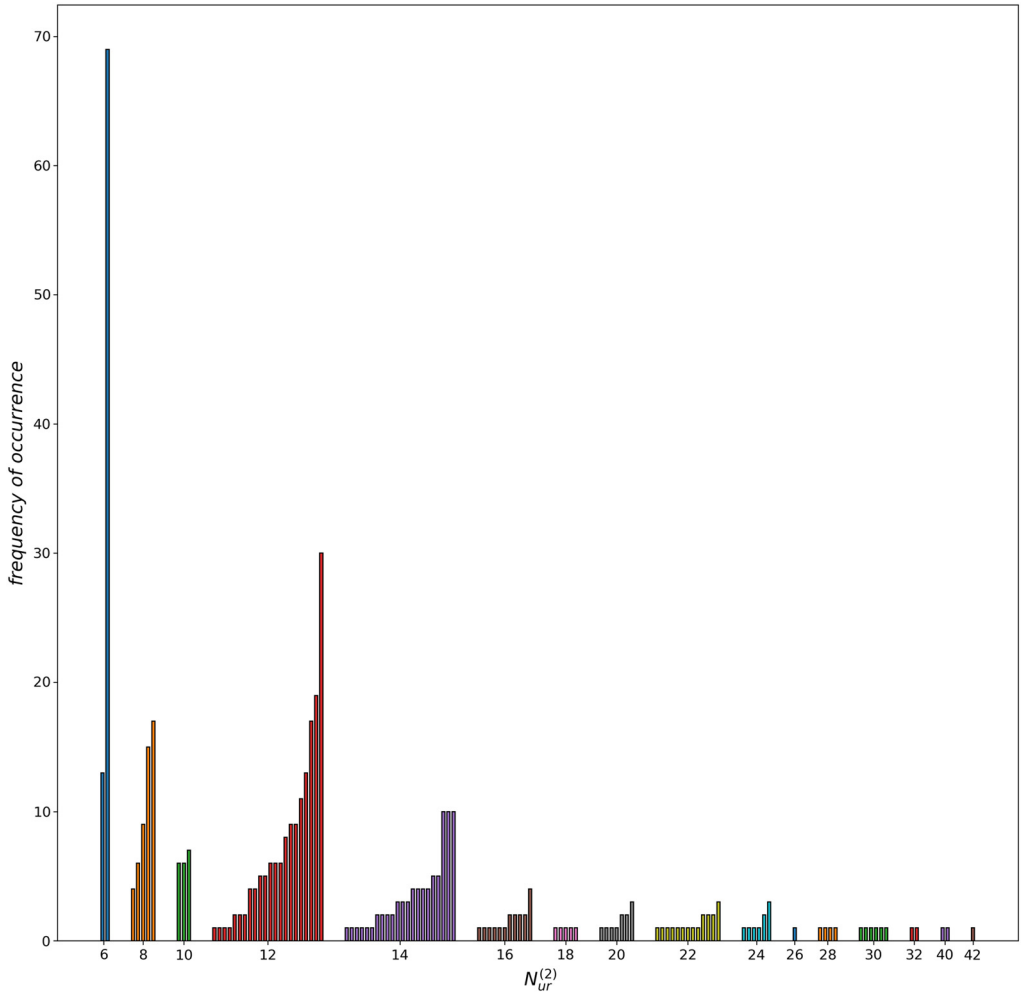


Fig. 6. Bar chart of MSSM-like $\mathbb{Z}_6$-II models M found using the *hidden* E_8 contrast pattern $N_{\rm ur}^{(2)}(M) \geq 6$. On the horizontal axis we give $N_{\rm ur}^{(2)}(M)$ corresponding to the number of unbroken roots from the hidden E_8. The vertical axis gives the corresponding frequency of occurrence. Each bar stands for an inequivalent MSSM-like model M. The number of copies of each model is shown by the height of the bar. As an example, take $N_{\rm ur}^{(2)}(M) = 10$, i.e. the green bars: There are three inequivalent MSSM-like models, each represented by one bar. These inequivalent models have 6, 6 and 7 copies in the whole dataset, respectively. Note that in this chart only those MSSM-like models appear that have G'_{SM} only in one E_8, since the notion of hidden E_8 is ambiguous otherwise.

to 136 in the *hidden* E_8 dataset. In the next section, we will investigate the reasons for this and present a solution that will lead to many new inequivalent MSSM-like models.

4.2.2. The dynamic hidden E_8 *contrast pattern*

Next, we analyze the effect of the *hidden* E_8 contrast pattern in more detail in order to identify a way to improve this constraint further. To do so, we take the (equivalent) MSSM-like $\mathbb{Z}_6$-II models from the *hidden* E_8 dataset and visualize how many MSSM-like models M appear for various values of $N_{\rm ur}^{(2)}(M)$, see Fig. 6. From this chart we see that the models with small

Table 5
Change of the growth rate for higher threshold values X of our contrast pattern $N_{\text{ur}}^{(2)}(M) \geq X$, where the reference point is $\text{gr}(N_{\text{ur}}^{(2)}(M) \geq 6) = 1$, since this analysis is done in the *hidden* E_8 dataset.

X	6	8	10	12	...	20	22	...	30	32	...	40	42
$\text{gr}(N_{\text{ur}}^{(2)}(M) \geq X)$	1	5	6	6	...	12	26	...	48	75	...	73	82

numbers of unbroken roots $N_{\text{ur}}^{(2)}(M) \in \{6, \ldots, 14\}$ are heavily oversampled in the *hidden* E_8 dataset, while it seems to be very difficult to construct models with $N_{\text{ur}}^{(2)}(M) \geq 30$. Moreover, note that especially for $N_{\text{ur}}^{(2)}(M) = 22$ the bar chart shows a lot of different bars, i.e. there are many inequivalent MSSM-like models with $N_{\text{ur}}^{(2)}(M) = 22$. This suggests that the diversity of inequivalent MSSM-like models may lie in some areas of the $\mathbb{Z}_6$-II orbifold landscape where models have larger hidden sector gauge groups. Furthermore, we investigate the change of the growth rate for higher threshold values X of our contrast pattern $N_{\text{ur}}^{(2)}(M) \geq X$ and obtain Table 5. Therefore, it seems very promising to change the threshold value $X = 6$ of the contrast pattern $N_{\text{ur}}^{(2)}(M) \geq X$ into a dynamic variable X. We call this new constraint *dynamic hidden* E_8. By applying this dynamic contrast pattern, we hope that the sampling among the various sizes of the hidden sector gets more balanced. Furthermore, we expect a boost in the number of MSSM-like models due to the increasing growth rate for higher values of $N_{\text{ur}}^{(2)}(M)$.

We perform an intensive search based on the *dynamic hidden* E_8 contrast pattern for various values of the threshold X and different constraints on the Wilson lines: First, we allow all Wilson lines to be non-trivial, then we turn off either $W_3 = W_4$ or W_5. As a result, we obtain a new dataset (which we also call *dynamic hidden* E_8), see Table 3. Compared to the *hidden* E_8 dataset with 136 inequivalent MSSM-like $\mathbb{Z}_6$-II models we now have in total 415 MSSM-like models. This is already more than in any existing $\mathbb{Z}_6$-II search [17–20]. Hence, we were able to significantly improve the search for inequivalent MSSM-like models in the $\mathbb{Z}_6$-II orbifold landscape. Moreover, this search solves the puzzle of the absence of MSSM-like models in the case $W_3 = W_4 = (0^{16})$: So far, it was not possible to find any MSSM-like model if the order 3 Wilson line is turned off, even though there is no theoretical obstruction for such a model to exist. Now, we have identified two MSSM-like $\mathbb{Z}_6$-II models with $W_3 = W_4 = (0^{16})$ as can be seen in Table 3. These models are equipped with a phenomenologically appealing $\Delta(54)$ flavor symmetry. Thus, we present these models in some detail in section 5.

A few remarks are in order. It is clear that previous searches based on the traditional approach as well as those presented in this paper are in general not exhaustive. During any random search process the number of inequivalent MSSM-like models will follow a saturation curve [44]. Consequently, the effort for creating a new inequivalent MSSM-like model growth exponentially during sampling. Thus, we believe that any attempt to reach our result using a basic random search would take an unrealizable amount of computing time and should be considered only a theoretical possibility rather than an alternative approach. So, why is our new search strategy so successful? Astonishingly, it turns out that a huge fraction of the diversity of MSSM-like models lies in areas of the heterotic orbifold landscape where the hidden sector gauge group is large, see Fig. 7. In more detail, using the *dynamic hidden* E_8 contrast pattern we could (i) obtain many new MSSM-like models with $N_{\text{ur}}^{(2)}(M) = X$ for $X \in \{34, 36, 44, 46, 56, 60, 62, 72, 74, 84\}$ and (ii) resolve the richness of MSSM-like models for higher X values, e.g. with $X \in \{30, 40, 42\}$. This was possible since, compared to Fig. 6, the new search strategy that led to Fig. 7 focuses especially on the regions with higher X values. These large hidden sector gauge groups can

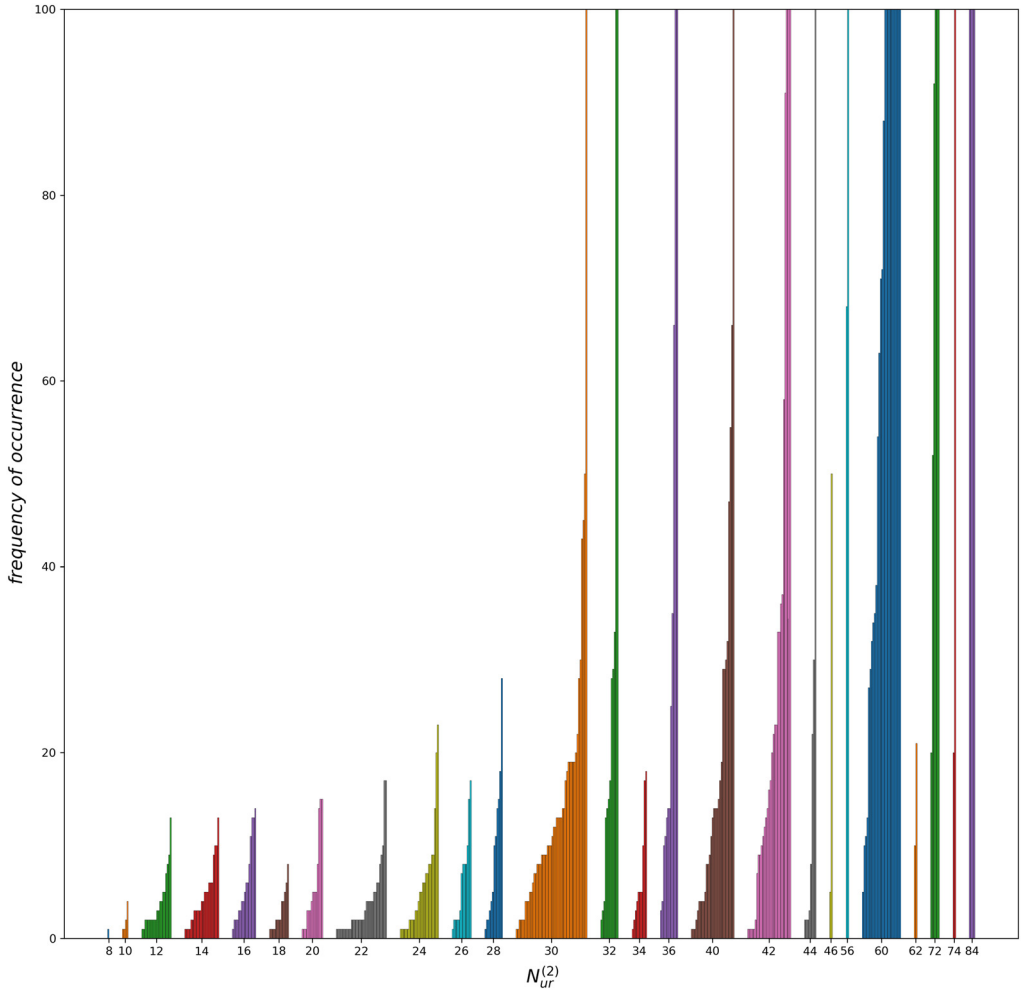


Fig. 7. Bar chart of MSSM-like $\mathbb{Z}_6$-II models M found using the *dynamic hidden* E_8 contrast pattern, cf. Fig. 6. Note that increasing the threshold value X of the contrast pattern $N_{\text{ur}}^{(2)}(M) \geq X$ leads to a deeper search in those areas of the $\mathbb{Z}_6$-II orbifold landscape that were insufficiently sampled by the static search using $X = 6$.

have direct physical implications related to supersymmetry breaking via gaugino condensation at rather high energies [35] and have to be studied in more detail.

4.2.3. *The* U-sector *contrast pattern*

On the basis of our *hidden* E_8 and *dynamic hidden* E_8 datasets we want to search for further contrast patterns. To do so, we follow the same logic as in section 4.2.1 and apply a decision tree on the remaining features $N_{\text{U}_a}^{(\alpha)}(M)$ to our new, combined dataset. For computational reasons we downsample our background of ~~MSSM~~-like models. This means we only work with a fraction of $\sim 50\%$ of the total dataset. This is a valid approach since we have so much data that the actual statistics for the decision tree will not change in a relevant way even if the whole dataset had been given. Furthermore, for the rare and important MSSM-like models we keep all inequivalent MSSM-like models, as described in section 4.2.1.

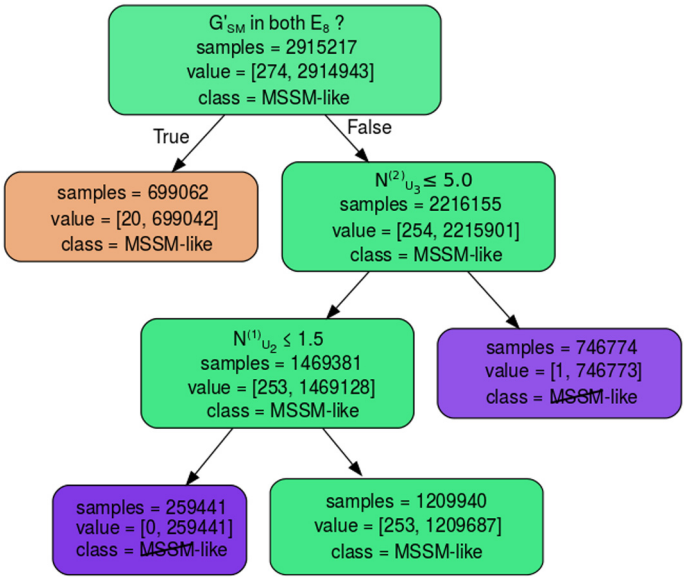


Fig. 8. Decision tree on the *hidden* E_8 and *dynamic hidden* E_8 datasets from Table 3, evaluated on the training set with $L_{12} = 10^5$, see eq. (24). We can extract the contrast pattern ($N^{(1)}_{U_2} \geq 2$ and $N^{(2)}_{U_3} \leq 5$) for MSSM-like models from this tree. One observes that this tree misclassifies one MSSM-like model after the $N^{(2)}_{U_3}$-split in order to get better performance. See also Fig. 5 for further details.

Then, the decision tree is trained with the same aim as before to classify all MSSM-like models correctly in both, training and validation set. However, it turns out that this is rather difficult due to two MSSM-like models: One of these models is misclassified during training, see Fig. 8, the other one during validation. It turns out that these two MSSM-like models are the special $\Delta(54)$ models, where $W_3 = W_4 = (0^{16})$. Due to the fact that these models lie in a very specific area within the $\mathbb{Z}_6$-II orbifold landscape, we decided to accept a misclassification of these models but with the benefit of obtaining a new contrast pattern that yields a further, significant reduction of the $\mathbb{Z}_6$-II orbifold landscape. Doing so, we identify a new contrast pattern

$$c_{\text{U-sector}} = \left\{ N^{(1)}_{U_2}(M) \geq 2\,,\; N^{(2)}_{U_3}(M) \leq 5 \right\}, \tag{28}$$

for a model M to have an increased probability to be MSSM-like. We call this contrast pattern *U-sector* as it gives bounds on the number of certain bulk matter fields, charged under the first or second E_8 factor, depending on $\alpha = 1, 2$, respectively. Using this new constraint on top of the previous ones, the estimated growth rate reads[7]

$$\text{gr}\left(c_{\text{U-sector}}, D^{N_{\text{ur}}}_{\text{MSSM-like}}, D^{N_{\text{ur}}}_{\cancel{\text{MSSM}}\text{-like}}\right) \approx \frac{0.999}{\frac{1\,209\,687}{2\,215\,901}} \approx 1.8 > 1\,, \tag{29}$$

where $D^{N_{\text{ur}}}$ is obtained by combining the datasets *hidden* E_8 and *dynamic hidden* E_8 from Table 3. Some remarks on the subtleties of the *U-sector* constraints are in order:

[7] Note that the estimated growth rate is computed based on the numbers of equivalent models. To do so, the same random split as for the training data of the decision tree has to be applied to the equivalent MSSM-like models from $D^{N_{\text{ur}}}$, yielding 8466 models. These numbers reduce to 8087 for models having G'_{SM} only in the first E_8 factor and, finally, to 8082 models fulfilling $c_{\text{U-sector}}$. Consequently, $\text{supp}(c_{\text{U-sector}}, D^{N_{\text{ur}}}_{\text{MSSM-like}}) = \frac{8082}{8087} \approx 0.999$.

gr(c): Contrary to the *hidden* E_8 contrast pattern, the *U-sector* contrast pattern can possibly exclude models which have $G'_{\rm SM}$ in both E_8 factors: in the case of $N^{(2)}_{\rm ur}(M) \geq 6$ models with $G'_{\rm SM}$ in both E_8 factors fulfill the even stronger condition $N^{(\alpha)}_{\rm ur}(M) \geq 8$ for $\alpha = 1, 2$. This can not be guaranteed for the *U-sector* constraint. Therefore, the growth rate in eq. (29) is estimated using only those models where $G'_{\rm SM}$ is exclusively in the first E_8 factor.

$N^{(1)}_{\mathrm{U}_2}$: Interestingly, the $\Delta(54)$ MSSM-like models excluded by the constraint $N^{(2)}_{\mathrm{U}_3}(M) \leq 5$ do obey the subsequent constraint $N^{(1)}_{\mathrm{U}_2}(M) \geq 2$ on the number of bulk matter from the U_2 sector and charged under the first E_8. Even though the decision tree decided strictly correct (by taking the statistics into account for optimization) and misclassified these two MSSM-like models, it is still appealing that all MSSM-like models (with $G'_{\rm SM} \in \mathrm{E}^{(1)}_8$) obey this constraint. This observation might be worth further investigations.

Implementing the *U-sector* contrast pattern into our search algorithm displayed in Fig. 1 and performing an intensive search (using all Wilson lines or turning off W_5 by hand), we obtain our final dataset, called *U-sector*, see Table 3. The results show once more the strength of the contrast data mining technique applied to the heterotic orbifold landscape: The probability to find MSSM-like models has increased further as shown in Fig. 9 and the *U-sector* contrast pattern generalizes to the $\mathbb{Z}_6$-II landscape such that we obtained many new inequivalent MSSM-like models. Starting from 395 inequivalent MSSM-like models in the *dynamic hidden* E_8 dataset we obtain now 459 models. Finally, combining all datasets yields in total 468 inequivalent MSSM-like $\mathbb{Z}_6$-II models.[8]

To summarize, we were able to significantly exceed all previous searches for MSSM-like $\mathbb{Z}_6$-II models [17–20] by excluding those regions in the $\mathbb{Z}_6$-II orbifold landscape where most likely no MSSM-like model exists. It is tempting to speculate that some of our contrast patterns might even be *necessary* conditions for *all* MSSM-like models. Moreover, we learned some general features of MSSM-like models that can be produced in the $\mathbb{Z}_6$-II orbifold landscape: We were able to identify constraints on physical quantities that can be interpreted and analyzed directly. Later, in section 6, we will show that the *hidden* E_8 contrast pattern can be transferred to other orbifold geometries while the lower bound of this constraint will be sensitive to the orbifold geometry under consideration.

5. $\mathbb{Z}_6$-II orbifold models with $\Delta(54)$ flavor symmetry

Out of the 468 inequivalent MSSM-like models based on the $\mathbb{Z}_6$-II orbifold geometry, there are two models with vanishing Wilson lines in the $\mathbb{Z}_3$ torus, i.e. $W_3 = W_4 = (0^{16})$. Hence, these MSSM-like models are equipped with a $\Delta(54)$ (R-) flavor symmetry [49–51], where localized matter fields transform in three-dimensional representations of $\Delta(54)$.

Both models are very similar. They are based on the same shift (shift No. 18 according to the enumeration of ref. [52]), but in different representations. This shift breaks the ten-dimensional $\mathrm{E}_8 \times \mathrm{E}_8$ gauge group to

$$\mathrm{SO}(10) \times \mathrm{SU}(3) \times \mathrm{U}(1) \quad \text{and} \quad \mathrm{SO}(12) \times \mathrm{SU}(2) \times \mathrm{U}(1)\,, \tag{30}$$

[8] Combining these 468 MSSM-like models with the known models from the literature yields 481 inequivalent MSSM-like models, see Table 8.

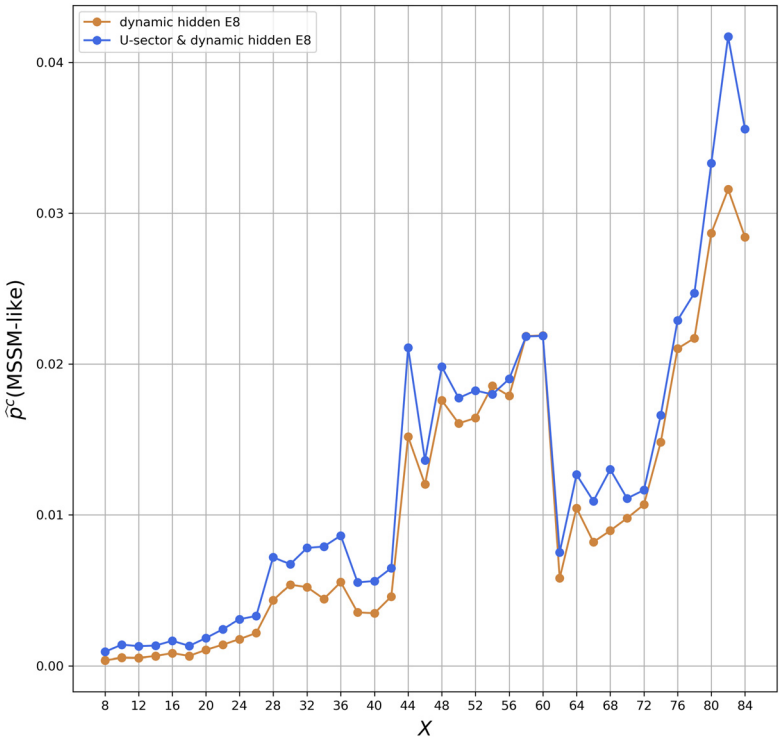


Fig. 9. Comparison of two different *dynamic hidden* E_8 searches: (i) without and (ii) with the additional *U-sector* constraint. For both cases, we display the probability $\hat{p}^c$(MSSM-like) (estimated on the sample) to find MSSM-like models under the constraint of the respective contrast pattern c as a function of the threshold value X, where c is (i) $N_{\text{ur}}^{(2)} \geq X$ and (ii) a combination of case (i) and the *U-sector* constraint, respectively.

in the first and second E_8, respectively. As a remark, this shift is different from the two local SO(10) shifts of ref. [19]. In a next step, SO(10) (and the hidden SO(12) $\times$ SU(2) gauge group) is broken by the Wilson lines W_5 and W_6 to the Standard Model gauge group, while the SU(3) factor remains unbroken as additional gauged $\text{SU}(3)_{\text{flavor}}$ flavor symmetry (Hence, the full flavor symmetry is actually $\text{SU}(3)_{\text{flavor}} \times \Delta(54)$). Consequently, both models share the same four-dimensional observable gauge group originating from the first E_8 factor, i.e.

$$\text{SU}(3)_{\text{flavor}} \times \text{SU}(3)_{\text{C}} \times \text{SU}(2)_{\text{L}} \times \text{U}(1)_Y \times \text{U}(1)^2 \times \text{hidden sector}\,. \tag{31}$$

Some details of these special MSSM-like $\mathbb{Z}_6$-II orbifold models are given in the following.

5.1. $\Delta(54)$ *MSSM #1 from the* $\mathbb{Z}_6$*-II orbifold*

The first $\mathbb{Z}_6$-II model with $\Delta(54)$ flavor symmetry is defined by the shift vector

$$V = \left(\tfrac{385}{12}, \tfrac{103}{12}, \tfrac{89}{12}, \tfrac{55}{12}, \tfrac{15}{4}, \tfrac{41}{12}, \tfrac{13}{4}, \tfrac{7}{12}\right), \left(\tfrac{145}{4}, \tfrac{139}{12}, \tfrac{33}{4}, \tfrac{25}{4}, \tfrac{47}{12}, \tfrac{11}{4}, \tfrac{7}{4}, -\tfrac{7}{4}\right)\,, \tag{32}$$

the Wilson lines $W_3 = W_4 = (0^{16})$, and

$$W_5 = \left(\tfrac{17}{2}, 5, 2, -1, 1, \tfrac{5}{2}, -1, -2\right), \left(5, 1, -1, 0, 1, \tfrac{1}{2}, 0, \tfrac{3}{2}\right)\,, \tag{33a}$$

Table 6
Massless matter spectrum of the first $\mathbb{Z}_6$-II model with $\Delta(54)$ flavor symmetry. Many SM matter fields build three-dimensional representations of $\Delta(54)$: For instance, all 6 quark-doublets q_i combine to two three-dimensional representations of $\Delta(54)$, similar for all partners $\bar{q}_i$, 6 out of 10 lepton-doublets (or down-Higgs) ℓ_i, and 3 out of 7 anti-lepton-doublets (or up-Higgs) $\bar{\ell}_i$. Furthermore, several SM singlets s_i^0 are three-dimensional representations of $\Delta(54)$. Finally, the SM singlets f_i and $\bar{f}_i$ are flavons of $\mathrm{SU}(3)_{\mathrm{flavor}}$ (interestingly, some of these flavons are simultaneously triplets of $\Delta(54)$).

#	Irrep	Labels	#	Irrep	Labels
6	$(\mathbf{1};\mathbf{3},\mathbf{2};\mathbf{1},\mathbf{1})_{-\frac{1}{6}}$	q_i	3	$(\mathbf{1};\bar{\mathbf{3}},\mathbf{2};\mathbf{1},\mathbf{1})_{\frac{1}{6}}$	$\bar{q}_i$
1	$(\mathbf{3};\bar{\mathbf{3}},\mathbf{1};\mathbf{1},\mathbf{1})_{\frac{2}{3}}$	$\bar{u}_i$			
1	$(\bar{\mathbf{3}};\bar{\mathbf{3}},\mathbf{1};\mathbf{1},\mathbf{1})_{-\frac{1}{3}}$	$\bar{d}_i$			
5	$(\mathbf{1};\bar{\mathbf{3}},\mathbf{1};\mathbf{1},\mathbf{1})_{-\frac{1}{3}}$	$\bar{d}_i$	5	$(\mathbf{1};\mathbf{3},\mathbf{1};\mathbf{1},\mathbf{1})_{\frac{1}{3}}$	d_i
10	$(\mathbf{1};\mathbf{1},\mathbf{2};\mathbf{1},\mathbf{1})_{\frac{1}{2}}$	ℓ_i	7	$(\mathbf{1};\mathbf{1},\mathbf{2};\mathbf{1},\mathbf{1})_{-\frac{1}{2}}$	$\bar{\ell}_i$
1	$(\mathbf{3};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{1})_{-1}$	$\bar{e}_1$			
12	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{1})_{\frac{1}{2}}$	s_i^+	12	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{1})_{-\frac{1}{2}}$	s_i^-
4	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\bar{\mathbf{3}})_{\frac{1}{2}}$	s_i^+	2	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{3})_{-\frac{1}{2}}$	s_i^-
2	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{3})_{\frac{1}{2}}$	s_i^+	4	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\bar{\mathbf{3}})_{-\frac{1}{2}}$	s_i^-
4	$(\mathbf{1};\mathbf{1},\mathbf{1};\bar{\mathbf{3}},\mathbf{1})_{\frac{1}{2}}$	s_i^+	2	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{3},\mathbf{1})_{-\frac{1}{2}}$	s_i^-
2	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{3},\mathbf{1})_{\frac{1}{2}}$	s_i^+	4	$(\mathbf{1};\mathbf{1},\mathbf{1};\bar{\mathbf{3}},\mathbf{1})_{-\frac{1}{2}}$	s_i^-
11	$(\bar{\mathbf{3}};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{1})_0$	f_i	10	$(\mathbf{3};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{1})_0$	$\bar{f}_i$
17	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{1})_0$	s_i^0			
7	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{3},\mathbf{1})_0$	s_i^0	6	$(\mathbf{1};\mathbf{1},\mathbf{1};\bar{\mathbf{3}},\mathbf{1})_0$	s_i^0
7	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\mathbf{3})_0$	s_i^0	6	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1},\bar{\mathbf{3}})_0$	s_i^0
1	$(\mathbf{1};\mathbf{1},\mathbf{1};\bar{\mathbf{3}},\mathbf{3})_0$	s_i^0	1	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{3},\bar{\mathbf{3}})_0$	s_i^0
1	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{3},\mathbf{3})_0$	s_i^0			

$$W_6 = \left(\tfrac{1}{2}, -\tfrac{3}{2}, -\tfrac{1}{2}, 3, 2, \tfrac{3}{2}, 0, 1\right), \left(2, -\tfrac{3}{2}, \tfrac{1}{2}, -1, 2, \tfrac{7}{2}, 0, -\tfrac{7}{2}\right) . \tag{33b}$$

It is called *MSSM431* in the model-file "Z6-II_1_1.txt" [36] that can be loaded to the `orbifolder`. For this model, the hidden gauge group is broken to

$$\mathrm{SU}(3) \times \mathrm{SU}(3) \times \mathrm{U}(1)^4 . \tag{34}$$

The massless string spectrum is summarized in Table 6. Interestingly, the spectrum contains flavons that are triplets under both $\Delta(54)$ and $\mathrm{SU}(3)_{\mathrm{flavor}}$. Hence, their vacuum expectation values could break the full flavor group to a diagonal subgroup.

5.2. $\Delta(54)$ *MSSM #2 from the* $\mathbb{Z}_6$*-II orbifold*

The second $\mathbb{Z}_6$-II model with $\Delta(54)$ flavor symmetry is given by the shift vector

$$V = \left(\tfrac{247}{4}, \tfrac{69}{4}, \tfrac{61}{4}, \tfrac{137}{12}, \tfrac{101}{12}, \tfrac{65}{12}, \tfrac{11}{4}, -\tfrac{1}{4}\right)\left(\tfrac{167}{3}, \tfrac{83}{6}, \tfrac{34}{3}, \tfrac{55}{6}, \tfrac{47}{6}, \tfrac{29}{6}, \tfrac{13}{6}, \tfrac{1}{6}\right) , \tag{35}$$

the Wilson lines $W_3 = W_4 = (0^{16})$, and

$$W_5 = \left(34, \tfrac{19}{2}, \tfrac{15}{2}, \tfrac{11}{2}, \tfrac{11}{2}, \tfrac{7}{2}, \tfrac{1}{2}, -3\right)\left(-\tfrac{113}{4}, -\tfrac{23}{4}, -\tfrac{17}{4}, -\tfrac{13}{4}, -\tfrac{27}{4}, -\tfrac{17}{4}, -\tfrac{3}{4}, \tfrac{5}{4}\right), \quad (36a)$$

$$W_6 = \left(-\tfrac{11}{4}, -\tfrac{1}{4}, \tfrac{5}{4}, -\tfrac{7}{4}, -\tfrac{11}{4}, \tfrac{1}{4}, -\tfrac{1}{4}, -\tfrac{11}{4}\right)\left(16, 1, 3, 2, 4, \tfrac{5}{2}, -\tfrac{1}{2}, -2\right). \quad (36b)$$

It is called *MSSM438* in the model-file "Z6-II_1_1.txt" [36]. In this case, the hidden gauge group is broken to

$$\mathrm{SU}(5) \times \mathrm{U}(1)^4\,, \quad (37)$$

in four dimensions and the massless string spectrum is given in Table 10 in appendix A. Compared to the $\Delta(54)$ MSSM #1, $\mathrm{SU}(3)_{\text{flavor}}$ and $\Delta(54)$ seem to have interchanged their role for many representations of quarks and leptons.

6. Geometry-dependent contrast patterns

The results from the previous sections were developed in the $\mathbb{Z}_6$-II (1, 1) orbifold geometry. However, the basic insights from this analysis can be transferred easily to other orbifold geometries. Foremost, the concept of a *hidden* E_8 contrast pattern can be applied directly to other orbifold geometries: The number of unbroken roots from the hidden E_8 is computed identically for all orbifold geometries and does not depend on some unknown sorting. Unfortunately, this is not given for the *U-sector* contrast pattern: The number of bulk matter fields $N^{(\alpha)}_{\mathrm{U}_a}$ for $a = 1, 2, 3$ depends on the twist vectors of a given orbifold geometry, see eq. (19). Then, the sorting of $N^{(\alpha)}_{\mathrm{U}_1}$, $N^{(\alpha)}_{\mathrm{U}_2}$ and $N^{(\alpha)}_{\mathrm{U}_3}$ is determined by the sorting of the entries in the twist vectors, which is typically sorted from small to large rotation angles. However, it is not clear why a particular U-sector might be special among the different sectors, i.e. if a special status of an U-sector is related to the sorting or to some nontrivial relation between all sectors.

Therefore, we begin with the *dynamic hidden* E_8 search and analyze the U-sector later. In order to apply the *dynamic hidden* E_8 contrast pattern to all $\mathbb{Z}_N$ orbifold geometries, we first have to identify the lower bound of $N^{(2)}_{\mathrm{ur}}(M)$ (being 6 in eq. (25)) for each $\mathbb{Z}_N$ orbifold geometry $\mathbb{O}$. Thus, we define

$$X_{\min}(\mathbb{O}) \;=\; \min\left(\left\{ N^{(2)}_{\mathrm{ur}}(M) \;\middle|\; M \in D_{(\mathbb{O}\text{ from }[17,18])} \right\}\right). \quad (38)$$

To compute these bounds, we use the traditional searches of refs. [17,18] as a background search and split the combined dataset into datasets $D_{(\mathbb{O}\text{ from }[17,18])}$ corresponding to the different $\mathbb{Z}_N$ orbifold geometries $\mathbb{O}$. Then, the conventional search in refs. [17,18] can be seen as a search with $N^{(2)}_{\mathrm{ur}} \geq 0$ in our approach. Thus, we can analyze the results of the traditional search [17,18] to obtain the lower bounds $X_{\min}(\mathbb{O})$ for all $\mathbb{Z}_N$ orbifold geometries. The results are stated in Table 7. Moreover, the background datasets $D_{(\mathbb{O}\text{ from }[17,18])}$ allow us to focus the *dynamic hidden* E_8 contrast pattern on thresholds greater than $X_{\min}(\mathbb{O})$, i.e. $N^{(2)}_{\mathrm{ur}} > X_{\min}(\mathbb{O})$, in order to save computational resources in our search.

First, let us state the results of our search in Table 8: For each $\mathbb{Z}_N$ orbifold geometry, we give the numbers of inequivalent MSSM-like orbifold models that we found using our *dynamic hidden* E_8 contrast pattern and compare these numbers to the literature. Several remarks are in order. One can observe that the *dynamic hidden* E_8 search was able to find many new inequivalent MSSM-like orbifold models in almost all orbifold geometries. Foremost, the different $\mathbb{Z}_6$-II orbifold geometries as well as the $\mathbb{Z}_{12}$-I case have improved strongly using our contrast patterns. Note that for $\mathbb{Z}_6$-II (1, 1), the 13 additional MSSM-like models from refs. [17,18] fulfill

Table 7
Lower bounds $X_{\min}(\mathbb{O})$ on the number $N_{\text{ur}}^{(2)}$ of unbroken roots from the hidden E_8 factor for MSSM-like orbifold models from refs. [17,18] for various $\mathbb{Z}_N$ orbifold geometries $\mathbb{O}$ (where "all" refers to the various lattices for a given $\mathbb{Z}_N$ orbifold geometry, as given in the first column of Table 8).

Orbifold geometry $\mathbb{O}$	$\mathbb{Z}_4$	$\mathbb{Z}_6$-I	$\mathbb{Z}_6$-II	$\mathbb{Z}_7$	$\mathbb{Z}_8$-I			$\mathbb{Z}_8$-II		$\mathbb{Z}_{12}$-I	$\mathbb{Z}_{12}$-II
	all	all	all	(1,1)	(1,1)	(2,1)	(3,1)	(1,1)	(2,1)	all	(1,1)
$X_{\min}(\mathbb{O})$	4	12	6	56	6	6	4	0	4	4	6

Table 8
Table of inequivalent MSSM-like orbifold models for all $\mathbb{Z}_N$ orbifold geometries, see also [36]. Note that the numbers of MSSM-like orbifold models listed in the third column differ from those in ref. [18]. This is due to an improvement of the `orbifolder` which has led to a better comparison of models and identified some duplicates in these sets. The last column, gives our final results: the numbers of inequivalent MSSM-like orbifold models obtained by merging the three datasets of the previous columns.

Inequivalent MSSM-like orbifold models					
Orbifold geometry		# MSSM-like from [17]	# MSSM-like from [18]	# MSSM-like using contrast patterns	# MSSM-like 'merged'
$\mathbb{Z}_4$	(2,1)	128	138	125	179
	(3,1)	25	26	33	33
$\mathbb{Z}_6$-I	(1,1)	31	30	31	31
	(2,1)	31	30	31	31
$\mathbb{Z}_6$-II	(1,1)	348	363	468	481
	(2,1)	338	349	395	443
	(3,1)	350	351	415	482
	(4,1)	334	354	407	464
$\mathbb{Z}_7$	(1,1)	0	1	1	1
$\mathbb{Z}_8$-I	(1,1)	263	256	248	271
	(2,1)	164	155	144	164
	(3,1)	387	377	408	430
$\mathbb{Z}_8$-II	(1,1)	638	1833	1259	2289
	(2,1)	260	489	349	555
$\mathbb{Z}_{12}$-I	(1,1)	365	556	610	625
	(2,1)	385	554	607	625
$\mathbb{Z}_{12}$-II	(1,1)	211	352	365	435

all our constraints derived in section 4. Consequently, even though these models were missed in our search, they are part of our search area. Hence, these models would have been found in an extended search. A great success of our contrast patterns is also given by the appearance of the MSSM-like $\mathbb{Z}_7$ model. This model was found so far only in refs. [18,39] using an orbifold-specific search strategy, as described in appendix A of ref. [39]. Also the $\mathbb{Z}_6$-I orbifold geometry is remarkable: In this case, we find a huge amount of equivalent MSSM-like models but only 31 inequivalent ones remain. These 31 inequivalent models were found very easily by searching in areas of the $\mathbb{Z}_6$-I orbifold landscape with large hidden sector gauge groups, cf. the lower bound $X_{\min}(\mathbb{Z}_6\text{-I}) = 12$ given in Table 7. Note that the lower bounds are computed for those models where G'_{SM} appears in one E_8 factor only. While most of the bounds in Table 7 are weaker than

Table 9
U-sector constraints for various $\mathbb{Z}_N$ orbifold geometries, based on the merged datasets of Table 8. We neglect the $\mathbb{Z}_7$ orbifold geometry because there is only one MSSM-like model available.

Orbifold geometry		$N_{U_1}^{(1)}$	$N_{U_2}^{(1)}$	$N_{U_3}^{(1)}$	$N_{U_1}^{(2)}$	$N_{U_2}^{(2)}$	$N_{U_3}^{(2)}$	gr(c)
$\mathbb{Z}_4$	(2,1)			≥ 4			≤ 1	5.32
	(3,1)			≥ 4			≤ 1	6.92
$\mathbb{Z}_6$-I	(1,1)	≥ 13		≥ 14				2.79
	(2,1)	≥ 13		≥ 14				2.78
$\mathbb{Z}_6$-II	(1,1)		≥ 2				≤ 5	1.81
	(2,1)		≥ 2				≤ 5	1.60
	(3,1)		≥ 2				≤ 5	1.70
	(4,1)		≥ 2				≤ 5	1.86
$\mathbb{Z}_8$-I	(1,1)			≥ 4			≤ 25	1.22
	(2,1)			≥ 4			≤ 25	1.23
	(3,1)			≥ 8				2.21
$\mathbb{Z}_8$-II	(1,1)			≥ 4			≤ 41	1.61
	(2,1)			≤ 3			≤ 1	1.78
				≥ 4			≤ 41	1.01
$\mathbb{Z}_{12}$-I	(1,1)	≤ 10		≥ 2				1.24
	(2,1)	≤ 10		≥ 2				1.24
$\mathbb{Z}_{12}$-II	(1,1)	≥ 2					≤ 5	1.71

$N_{ur}^{(1)} \geq 8$ from section 3.2, we have to be careful in the cases of both $\mathbb{Z}_6$-I orbifold geometries. There, we find a lower bound $X_{min}(\mathbb{Z}_6\text{-I}) = 12$. Thus, our search algorithm could in principle miss MSSM-like $\mathbb{Z}_6$-I models where each E_8 factor contains G'_{SM}. We analyze these problematic models separately and find a lower bound $X_{min}(\mathbb{Z}_6\text{-I}) = 10$ for these cases (because $(N_{ur}^{(1)}, N_{ur}^{(2)})$ takes only the values $(8, 12)$ and $(10, 10)$ in the background dataset). This means that MSSM-like $\mathbb{Z}_6$-I models which have G'_{SM} in both E_8 factors are contained in our search $N_{ur}^{(2)} \geq 10$ for both $\mathbb{Z}_6$-I orbifold geometries.

Furthermore, it seems that for some orbifold geometries like $\mathbb{Z}_8$-II $(1, 1)$ the conventional approach has some advantages. However, a comparison is difficult since it is not known how much computational power was invested to obtain these numbers. Moreover, for $\mathbb{Z}_8$-II $(1, 1)$ our contrast patterns seem to be less efficient since there is no lower bound for $N_{ur}^{(2)}$, i.e. MSSM-like models with $N_{ur}^{(2)} = 0$ exist for $\mathbb{Z}_8$-II $(1, 1)$, and there are many inequivalent MSSM-like models for low values of $N_{ur}^{(2)}$, which can be reached by the conventional search algorithm as well. Hence, on first sight it seems that our search algorithm is too complex for such geometries and the additional effort in computing constraints is not rewarded. However, this conclusion is premature: The merged datasets in Table 8 show that our contrast patterns could still significantly improve the numbers of inequivalent MSSM-like orbifold models in these geometries. Thus, our search algorithm was able to find new MSSM-like orbifold models in corners of the landscape that were missed by the conventional approach.

On the basis of the 'merged' datasets we can now use decision trees to derive the *U-sector* constraints as in the case of the $\mathbb{Z}_6$-II orbifold geometry. The results are given in Table 9. Let us analyze the resulting *U-sector* contrast patterns in some detail: In nearly all orbifold geometries

it is possible to get a recall in the validation set of 1.00 and no MSSM-like model is missed in the training set. Only for the $\mathbb{Z}_6$-II orbifold geometries (1, 1), (2, 1) and (3, 1) a very few *false negative* predictions were made either in the validation set or in the training set. Another special case is the $\mathbb{Z}_8$-II (2, 1) orbifold geometry: In order to get a growth rate larger than one the decision tree had to split the set of MSSM-like models into two sets at the first node with a constraint on $N^{(1)}_{U_3}$. Then, for both sets a second split takes $N^{(2)}_{U_3}$ into account, resulting in a growth rate of 1.78 for the first set (containing only 3% of the MSSM-like $\mathbb{Z}_8$-II (2, 1) models from the training set) and a growth rate of 1.01 for the second set (containing 97% of the models), respectively. In this context, let us mention that for $\mathbb{Z}_8$-I (1, 1) it is possible to create another decision tree with a constraint $c'_{\mathrm{U-sector}} = \left\{N^{(1)}_{U_3} \geq 8, N^{(2)}_{U_3} \leq 25\right\}$ that has a recall value of 1.00 and $\mathrm{gr}(c'_{\mathrm{U-sector}}) = 2.18$, however, with the trade-off of missing one MSSM-like model from the training data.

Interestingly, it seems that the *U-sector* constraints in Table 9 show some patterns on their own. Foremost, it is remarkable that for a given twist vector the exact orbifold geometry (i.e. the choice of the six-torus which is enumerated by the number $l = 1, 2, \ldots$ in the label $(l, 1)$) does not have any significant effect. This could be used to extrapolate from one orbifold geometry to another. If the constraints are different within one twist vector one should be careful and probably take the weakest constraint, e.g. $N^{(3)}_{U_1} \geq 4$ for $\mathbb{Z}_8$-I. This can be seen as regularizing the machine learning model. One can use this insight to avoid overfitting and to use more statistics from other orbifold geometries. Additionally, one can observe that the hidden sector is completely dominated by a '$\leq$' constraint in the U_3-*sectors*, while the visible sector favors '$\geq$' (except for $\mathbb{Z}_{12}$-I and $\mathbb{Z}_6$-I). This shows that there is still more structure to explore in the heterotic orbifold landscape and, more importantly, that we are on the right track to obtain necessary conditions on both the observable E_8 factor, containing the MSSM, and the hidden E_8 factor.

7. Conclusion

In this paper, we have developed an advanced search strategy for MSSM-like orbifold models using the $\mathbb{Z}_6$-II (1, 1) orbifold geometry as a test case. We obtained a significant improvement from 363 inequivalent MSSM-like models [18] to 481, see Table 8. To do so, we used a technique called contrast data mining, where one identifies so-called contrast patterns that help to distinguish between MSSM-like models and others. In principle, this technique is easy to generalize to all orbifold geometries and, presumably, to other string compactifications. As a first step towards this, we analyzed all $\mathbb{Z}_N$ orbifold geometries in section 6 and showed that in all cases our contrast patterns significantly enhance the known datasets of MSSM-like orbifold models, see Table 8. Let us stress that this new search strategy is superior by orders of magnitudes with respect to the computing time. Theoretically, the conventional search algorithm can find *all* MSSM-like orbifold models. However, this would correspond to an unfeasible amount of computing time because the effort for finding a new MSSM-like model grows exponentially with the number of already constructed models. This fact was studied in detail in ref. [44] and can also be inferred from Figs. 3, 6 and 7. These figures show that the towers of already known orbifold models dominate the search and statistically keep growing before new orbifold models are expected to appear. Hence, with increasing search time the probability to find a new orbifold model is suppressed further and further.

Consequently, we believe that contrast patterns can be of great importance when studying the string landscape. In addition, contrast patterns are particularly useful as they have a clear

physical interpretation. In our setup of heterotic orbifolds, we identified the following contrast patterns: the number of unbroken roots in the hidden E_8 factor and the numbers of various bulk matter fields, charged under first or second E_8 factor. Hence, our contrast patterns are related to bulk fields that originate from the compactification of the ten-dimensional $E_8 \times E_8$ gauge bosons. Moreover, our contrast patterns have direct phenomenological implications, as they are important for supersymmetry breaking via hidden sector gaugino condensation [9,35], gauge-Higgs unification [53,54] and gauge-top unification [55]. Further studies along these lines have to follow.

Moreover, using the approach with contrast patterns it was possible to solve some long standing issues in the heterotic orbifold landscape, namely:

- We found many new MSSM-like orbifold models, especially in corners of the heterotic orbifold landscape that were hardly accessible by the conventional search algorithms, see Table 8.
- As stated in Table 3, it was possible to prove the existence of MSSM-like $\mathbb{Z}_6$-II models with vanishing Wilson line of order three, i.e. $W_3 = W_4 = (0^{16})$. This is the first time that such models are described in the literature. They might be phenomenologically interesting as they are equipped with a $\Delta(54)$ flavor symmetry, see section 5.
- Furthermore, using the new technique we were able to reproduce the only known MSSM-like model in the $\mathbb{Z}_7$ orbifold geometry [18,39]. This model could not be found by any random search so far. Instead, it was found using a method (described in appendix A of ref. [39]) that is not feasible for most other orbifold geometries.
- Moreover, even though the main aim of our search algorithm is to find *inequivalent* MSSM-like orbifold models, we obtain in addition an important byproduct: our contrast patterns significantly increase the probability to find MSSM-like orbifold models in certain regions of the heterotic orbifold landscape. In future applications, the models that are classified as equivalent by the `orbifolder` may differ in some other aspects, e.g. in their Yukawa couplings, see section 2.2. As soon as a preferred MSSM-like orbifold model is identified, our search algorithm allows to explore a specific part of the landscape in order to find models that have similar spectra but are not necessarily equivalent with respect to the full model.
- Also this work can be seen to be a fundamental step in applying further machine learning techniques to the heterotic orbifold landscape. In this paper, we are fighting the imbalance of the string theory datasets, i.e. we are trying to get enough data of the minority class, which is build up by the MSSM-like models. Especially, deep learning techniques need further research to handle unbalanced data, while traditional machine learning methods, i.e. non-neural-networks, are studied in detail in the context of imbalanced classification, see e.g. [56].

Finally, we want to state some preferred properties of contrast pattern that should be kept in mind when constructing new features in the future. These properties are useful for implementation as well as for the impact of a new contrast pattern. The most important property is that a new feature can be checked quickly and easily, since it will be computed several times during the successive search algorithm. In addition, it is an advantage if a contrast pattern is testable at each step of the successive construction, see Fig. 1. Therefore, a new feature must be a monotonically decreasing (increasing) function with respect to the successive creation of shifts and Wilson lines. Moreover, in combination with the monotonic behavior the constraint has to be a lower (upper) bound on the model. For example, the minimal number of unbroken roots is a good

contrast pattern since it can only decrease at each step in Fig. 1 and, therefore, it is a monotonically decreasing function with a lower bound. On the other side, the maximal number of bulk fields in a certain *U-sector* can only be checked at the last step in Fig. 1 since this number is also a monotonically decreasing function and any subsequently chosen Wilson line can decrease this value further. Hence, even though the number of bulk fields is a monotonically decreasing function with respect to the successive creation of shifts and Wilson lines, the *U-sector* contrast pattern is given by an upper bound, which weakens this contrast pattern.

In conclusion, this paper shows that techniques from data mining and machine learning can be applied successfully to the heterotic orbifold landscape and produce practical results, i.e. novel MSSM-like models that were out of reach using all traditional approaches so far. Further investigations in this direction have to be done in order to complete the set of contrast patterns. One might speculate that contrast patterns could ultimately help to identify an analytic formula for the construction of MSSM-like models in the heterotic orbifold landscape.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (SFB1258). We would like to thank Saúl Ramos-Sánchez and Hans Peter Nilles for useful discussions.

Appendix A. Spectrum of $\mathbb{Z}_6$-II MSSM # 2 with $\Delta(54)$ flavor symmetry

Table 10
Massless matter spectrum of the second $\mathbb{Z}_6$-II model with $\Delta(54)$ flavor symmetry. Many SM matter fields build three-dimensional representations of $\Delta(54)$: For instance, all 6 up-quarks $\bar{u}_i$ combine to two three-dimensional representations of $\Delta(54)$. Furthermore, several SM singlets s_i^0 are three-dimensional representations of $\Delta(54)$. Finally, the SM singlets f_i and $\bar{f}_i$ are flavons of $\mathrm{SU}(3)_{\mathrm{flavor}}$ (interestingly, some of these flavons are simultaneously triplets of $\Delta(54)$).

#	Irrep	Labels	#	Irrep	Labels
1	$(\mathbf{3};\bar{\mathbf{3}},\mathbf{2};\mathbf{1})_{\frac{1}{6}}$	q_i			
6	$(\mathbf{1};\mathbf{3},\mathbf{1};\mathbf{1})_{-\frac{2}{3}}$	$\bar{u}_i$	3	$(\mathbf{1};\bar{\mathbf{3}},\mathbf{1};\mathbf{1})_{\frac{2}{3}}$	u_i
1	$(\bar{\mathbf{3}};\mathbf{3},\mathbf{1};\mathbf{1})_{\frac{1}{3}}$	$\bar{d}_i$			
4	$(\mathbf{1};\mathbf{3},\mathbf{1};\mathbf{1})_{\frac{1}{3}}$	$\bar{d}_i$	4	$(\mathbf{1};\bar{\mathbf{3}},\mathbf{1};\mathbf{1})_{-\frac{1}{3}}$	d_i
11	$(\mathbf{1};\mathbf{1},\mathbf{2};\mathbf{1})_{-\frac{1}{2}}$	ℓ_i	8	$(\mathbf{1};\mathbf{1},\mathbf{2};\mathbf{1})_{\frac{1}{2}}$	$\bar{\ell}_i$
6	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1})_{1}$	$\bar{e}_i$	3	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1})_{-1}$	e_i
3	$(\mathbf{1};\bar{\mathbf{3}},\mathbf{1};\mathbf{1})_{\frac{1}{6}}$	$\bar{v}_i$	3	$(\mathbf{1};\mathbf{3},\mathbf{1};\mathbf{1})_{-\frac{1}{6}}$	v_i
12	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1})_{\frac{1}{2}}$	s_i^+	3	$(\bar{\mathbf{3}};\mathbf{1},\mathbf{1};\mathbf{1})_{-\frac{1}{2}}$	s_i^-
2	$(\bar{\mathbf{3}};\mathbf{1},\mathbf{1};\mathbf{1})_{\frac{1}{2}}$	s_i^+	3	$(\mathbf{3};\mathbf{1},\mathbf{1};\mathbf{1})_{-\frac{1}{2}}$	s_i^i
15	$(\mathbf{1};\mathbf{1},\mathbf{2};\mathbf{1})_{0}$	m_i			
2	$(\mathbf{1};\mathbf{1},\mathbf{2};\mathbf{5})_{0}$	m_i	1	$(\mathbf{1};\mathbf{1},\mathbf{2};\bar{\mathbf{5}})_{0}$	m_i
25	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{1})_{0}$	s_i^0			
11	$(\bar{\mathbf{3}};\mathbf{1},\mathbf{1};\mathbf{1})_{0}$	$\bar{f}_i$	10	$(\mathbf{3};\mathbf{1},\mathbf{1};\mathbf{1})_{0}$	f_i
7	$(\mathbf{1};\mathbf{1},\mathbf{1};\bar{\mathbf{5}})_{0}$	s_i^0	6	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{5})_{0}$	s_i^0
2	$(\mathbf{1};\mathbf{1},\mathbf{1};\bar{\mathbf{10}})_{0}$	s_i^0	1	$(\mathbf{1};\mathbf{1},\mathbf{1};\mathbf{10})_{0}$	s_i^0

References

[1] W. Lerche, D. Lüst, A.N. Schellekens, Nucl. Phys. B 287 (1987) 477.
[2] M.R. Douglas, J. High Energy Phys. 05 (2003) 046, arXiv:hep-th/0303194.
[3] L.J. Dixon, J.A. Harvey, C. Vafa, E. Witten, Nucl. Phys. B 261 (1985) 678.
[4] L.J. Dixon, J.A. Harvey, C. Vafa, E. Witten, Nucl. Phys. B 274 (1986) 285.
[5] A.E. Faraggi, Nucl. Phys. B 387 (1992) 239, arXiv:hep-th/9208024 [hep-th].
[6] T.P.T. Dijkstra, L.R. Huiszoon, A.N. Schellekens, Nucl. Phys. B 710 (2005) 3, arXiv:hep-th/0411129 [hep-th].
[7] V. Braun, Y.-H. He, B.A. Ovrut, T. Pantev, Phys. Lett. B 618 (2005) 252, arXiv:hep-th/0501070 [hep-th].
[8] F. Gmeiner, R. Blumenhagen, G. Honecker, D. Lüst, T. Weigand, J. High Energy Phys. 01 (2006) 004, arXiv:hep-th/0510170 [hep-th].
[9] K.R. Dienes, Phys. Rev. D 73 (2006) 106010, arXiv:hep-th/0602286 [hep-th].
[10] R. Blumenhagen, B. Körs, D. Lüst, S. Stieberger, Phys. Rep. 445 (2007) 1, arXiv:hep-th/0610327 [hep-th].
[11] L.B. Anderson, J. Gray, A. Lukas, E. Palti, Phys. Rev. D 84 (2011) 106005, arXiv:1106.4804 [hep-th].
[12] L.B. Anderson, J. Gray, A. Lukas, E. Palti, J. High Energy Phys. 06 (2012) 113, arXiv:1202.1757 [hep-th].
[13] M. Cvetič, D. Klevers, D.K. Mayorga Peña, P.-K. Oehlmann, J. Reuter, J. High Energy Phys. 08 (2015) 087, arXiv:1503.02068 [hep-th].
[14] M. Cvetič, L. Lin, M. Liu, P.-K. Oehlmann, J. High Energy Phys. 09 (2018) 089, arXiv:1807.01320 [hep-th].
[15] M. Cvetič, J. Halverson, L. Lin, M. Liu, J. Tian, Phys. Rev. Lett. 123 (10) (2019) 101601, arXiv:1903.00009 [hep-th].
[16] S. Groot Nibbelink, O. Loukas, J. High Energy Phys. 12 (2013) 044, arXiv:1308.5145 [hep-th].
[17] H.P. Nilles, P.K.S. Vaudrevange, Mod. Phys. Lett. A 30 (10) (2015) 1530008, arXiv:1403.1597 [hep-th].
[18] Y. Olguín-Trejo, R. Pérez-Martínez, S. Ramos-Sánchez, Phys. Rev. D 98 (10) (2018) 106020, arXiv:1808.06622 [hep-th].
[19] O. Lebedev, H.P. Nilles, S. Raby, S. Ramos-Sánchez, M. Ratz, P.K.S. Vaudrevange, A. Wingerter, Phys. Lett. B 645 (2007) 88, arXiv:hep-th/0611095 [hep-th].
[20] O. Lebedev, H.P. Nilles, S. Ramos-Sánchez, M. Ratz, P.K.S. Vaudrevange, Phys. Lett. B 668 (2008) 331, arXiv:0807.4384 [hep-th].
[21] D.K. Mayorga Peña, H.P. Nilles, P.-K. Oehlmann, J. High Energy Phys. 12 (2012) 024, arXiv:1209.6041 [hep-th].
[22] Y.-H. He, arXiv:1706.02714 [hep-th].
[23] D. Krefl, R.-K. Seong, Phys. Rev. D 96 (6) (2017) 066014, arXiv:1706.03346 [hep-th].
[24] F. Ruehle, J. High Energy Phys. 08 (2017) 038, arXiv:1706.07024 [hep-th].
[25] J. Carifio, J. Halverson, D. Krioukov, B.D. Nelson, J. High Energy Phys. 09 (2017) 157, arXiv:1707.00655 [hep-th].
[26] Y.-N. Wang, Z. Zhang, J. High Energy Phys. 08 (2018) 009, arXiv:1804.07296 [hep-th].
[27] K. Bull, Y.-H. He, V. Jejjala, C. Mishra, Phys. Lett. B 785 (2018) 65, arXiv:1806.03121 [hep-th].
[28] D. Klaewer, L. Schlechter, Phys. Lett. B 789 (2019) 438, arXiv:1809.02547 [hep-th].
[29] J. Halverson, B. Nelson, F. Ruehle, J. High Energy Phys. 06 (2019) 003, arXiv:1903.11616 [hep-th].
[30] A. Cole, G. Shiu, J. High Energy Phys. 03 (2019) 054, arXiv:1812.06960 [hep-th].
[31] A. Cole, A. Schachner, G. Shiu, J. High Energy Phys. 11 (2019) 045, arXiv:1907.10072 [hep-th].
[32] K. Bull, Y.-H. He, V. Jejjala, C. Mishra, Phys. Lett. B 795 (2019) 700, arXiv:1903.03113 [hep-th].
[33] A. Ashmore, Y.-H. He, B. Ovrut, arXiv:1910.08605 [hep-th].
[34] A. Mütter, E. Parr, P.K.S. Vaudrevange, Nucl. Phys. B 940 (2019) 113, arXiv:1811.05993 [hep-th].
[35] O. Lebedev, H.-P. Nilles, S. Raby, S. Ramos-Sánchez, M. Ratz, P.K.S. Vaudrevange, A. Wingerter, Phys. Rev. Lett. 98 (2007) 181602, arXiv:hep-th/0611203 [hep-th].
[36] E. Parr, P.K.S. Vaudrevange, The model-files for the orbifolder, which contain the gauge embeddings of all MSSM-like $\mathbb{Z}_N$ orbifold models, can be found as arXiv ancillary files of this paper, 2019.
[37] M. Fischer, M. Ratz, J. Torrado, P.K.S. Vaudrevange, J. High Energy Phys. 01 (2013) 084, arXiv:1209.3906 [hep-th].
[38] L.E. Ibáñez, H.P. Nilles, F. Quevedo, Phys. Lett. B 187 (1987) 25.
[39] S. Ramos-Sánchez, Fortschr. Phys. 10 (2009) 907, arXiv:0812.3560 [hep-th], Ph.D. Thesis (Advisor: H.P. Nilles).
[40] F. Plöger, S. Ramos-Sánchez, M. Ratz, P.K.S. Vaudrevange, J. High Energy Phys. 04 (2007) 063, arXiv:hep-th/0702176 [hep-th].
[41] H.P. Nilles, S. Ramos-Sánchez, P.K.S. Vaudrevange, A. Wingerter, Comput. Phys. Commun. 183 (2012) 1363, arXiv:1110.5229 [hep-th], http://projects.hepforge.org/orbifolder/.
[42] S. Ramos-Sánchez, P.K.S. Vaudrevange, J. High Energy Phys. 01 (2019) 055, arXiv:1811.00580 [hep-th].
[43] J. Fuchs, C. Schweigert, Symmetries, Lie Algebras and Representations: A Graduate Course for Physicists, University Press, Cambridge, UK, 1997, 438 p.

[44] K.R. Dienes, M. Lennek, Phys. Rev. D 75 (2007) 026008, arXiv:hep-th/0610319.
[45] G. Dong, J. Bailey, Contrast Data Mining: Concepts, Algorithms, and Applications, 1st ed., Chapman & Hall/CRC, 2012.
[46] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D.C. Gemperline, T. Augspurger, Y. Halchenko, J.B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M.L. Williams, C. Evans, C. Fitzgerald Brian, C. Fonnesbeck, A. Lee, A. Qalieh, mwaskom/seaborn: v0.8.1 (September 2017), September 2017.
[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, J. Mach. Learn. Res. 12 (2011) 2825.
[48] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.
[49] T. Kobayashi, H.P. Nilles, F. Plöger, S. Raby, M. Ratz, Nucl. Phys. B 768 (2007) 135, arXiv:hep-ph/0611020 [hep-ph].
[50] H.P. Nilles, M. Ratz, P.K.S. Vaudrevange, Fortschr. Phys. 61 (2013) 493, arXiv:1204.2206 [hep-ph].
[51] A. Baur, H.P. Nilles, A. Trautner, P.K.S. Vaudrevange, Nucl. Phys. B 947 (2019) 114737, arXiv:1908.00805 [hep-th].
[52] Y. Katsuki, Y. Kawamura, T. Kobayashi, N. Ohtsubo, Y. Ono, K. Tanioka, DPKU-8904, 1989.
[53] L.J. Hall, Y. Nomura, D. Tucker-Smith, Nucl. Phys. B 639 (2002) 307, arXiv:hep-ph/0107331 [hep-ph].
[54] M. Kubo, C.S. Lim, H. Yamashita, Mod. Phys. Lett. A 17 (2002) 2249, arXiv:hep-ph/0111327 [hep-ph].
[55] P. Hosteins, R. Kappl, M. Ratz, K. Schmidt-Hoberg, J. High Energy Phys. 07 (2009) 029, arXiv:0905.3323 [hep-ph].
[56] J.M. Johnson, T.M. Khoshgoftaar, J. Big Data 6 (1) (2019) 27.