Machine learning regularization for the minimum volume formula of toric Calabi-Yau 3-folds

Eugene Choi[1,*] and Rak-Kyeong Seong[1,2,†]

[1]*Department of Mathematical Sciences, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulsan 44919, South Korea*

[2]*Department of Physics, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulsan 44919, South Korea*

(Received 31 October 2023; accepted 31 January 2024; published 23 February 2024)

We present a collection of explicit formulas for the minimum volume of Sasaki-Einstein 5-manifolds. The cone over these 5-manifolds is a toric Calabi-Yau 3-fold. These toric Calabi-Yau 3-folds are associated with an infinite class of $4d$ $\mathcal{N}=1$ supersymmetric gauge theories, which are realized as world volume theories of D3-branes probing the toric Calabi-Yau 3-folds. Under the AdS/CFT correspondence, the minimum volume of the Sasaki-Einstein base is inversely proportional to the central charge of the corresponding $4d$ $\mathcal{N}=1$ superconformal field theories. The presented formulas for the minimum volume are in terms of geometric invariants of the toric Calabi-Yau 3-folds. These explicit results are derived by implementing machine learning regularization techniques that advance beyond previous applications of machine learning for determining the minimum volume. Moreover, the use of machine learning regularization allows us to present interpretable and explainable formulas for the minimum volume. Our work confirms that, even for extensive sets of toric Calabi-Yau 3-folds, the proposed formulas approximate the minimum volume with remarkable accuracy.

DOI: 10.1103/PhysRevD.109.046015

I. INTRODUCTION

Since the introduction of machine learning techniques in [1–13] for studying problems that occur in the context of string theory, machine learning—both supervised [14–20] and unsupervised [21–24]—has led to a variety of applications in string theory. A problem that appeared particularly suited for machine learning in 2017 [2] was the problem of identifying a formula for the minimum volume of Sasaki-Einstein 5-manifolds [25,26]. The cone over these Sasaki-Einstein 5-manifolds is a toric Calabi-Yau 3-fold [27,28]. Given that there are infinitely many toric Calabi-Yau 3-folds with corresponding Sasaki-Einstein 5-manifolds and that there is an infinite class of $4d$ $\mathcal{N}=1$ supersymmetric gauge theories associated to them via string theory [29–36], this beautiful correspondence between geometry and gauge theory was identified in [2] as an ideal test bed for introducing machine learning for string theory.

These $4d$ $\mathcal{N}=1$ supersymmetric gauge theories corresponding to toric Calabi-Yau 3-folds are realized as world-volume theories of D3-branes probing the Calabi-Yau singularities. Via the AdS/CFT correspondence [37–39], the minimum volume of the Sasaki-Einstein 5-manifolds is related to the maximized a-function [40–42] that gives the central charges of the corresponding $4d$ $\mathcal{N}=1$ superconformal field theories [43,44]. The proposal in [2] was that machine learning techniques can be used to give a formula of the minimum volume in terms of features taken from the toric diagram of the corresponding toric Calabi-Yau 3-folds. Such a formula would significantly simplify the computation of the minimum volume, which conventionally is computed by minimizing the volume function obtained from the equivariant index [25,26] or Hilbert series of the toric Calabi-Yau 3-fold [45–47].

In [2], we made use of multiple linear regression [48–52] and a combination of a regression model and a convolutional neural network (CNN) [53–56] to learn the minimum volume for toric Calabi-Yau 3-folds. As it is often the case for supervised machine learning [57,58], the models lacked interpretability and explainability, achieving high accuracies in estimating the minimum volume with giving only little insight into the mathematical structure and physical origin of the estimating formula.

In this work, we aim to highlight the pivotal role of regularization techniques in machine learning [58,59].

*xeugenechoi@gmail.com
†seong@unist.ac.kr

We demonstrate that employing regularized machine learning models can effectively address the limitations inherent in supervised machine learning, especially for problems that appear in string theory and, more broadly, for problems at the intersection of mathematics and physics. While the primary objective of regularization in machine learning is to prevent overfitting, certain versions of it can be employed to eliminate model parameters, echoing the spirit of regularization in quantum field theory.

By focusing on least absolute shrinkage and selection operator (Lasso) regularization [60] for polynomial and logarithmic regression models, we identify several candidate formulas for the minimum volume of Sasaki-Einstein 5-manifolds corresponding to toric Calabi-Yau 3-folds. The discovered formulas depend either on three or six parameters that come from features of the corresponding toric diagrams [27,28]; convex lattice polygons on $\mathbb{Z}^2$ that characterize uniquely the associated toric Calabi-Yau 3-fold. Compared to the extremely large number of parameters in the regression and CNN models used in our previous work in [2], the formulas obtained in this study are both presentable, interpretable, and most importantly reusable for the computation of the minimum volume for toric Calabi-Yau 3-folds.

II. CALABI-YAU 3-FOLDS AND QUIVER GAUGE THEORIES

In this work, we concentrate on noncompact toric Calabi-Yau 3-folds $\mathcal{X}$. These geometries can be considered as cones over Sasaki-Einstein 5-manifolds Y_5 [37–39,61–64]. The toric Calabi-Yau 3-folds are fully characterized by convex lattice polygons Δ on $\mathbb{Z}^2$ known as toric diagrams [27,28]. The associated Calabi-Yau singularities can be probed by D3-branes whose worldvolume theories form a class of $4d$ $\mathcal{N}=1$ supersymmetric gauge theories [29–36].

This class of $4d$ $\mathcal{N}=1$ supersymmetric gauge theories can be represented in terms of a T-dual type IIB brane configuration known as a brane tiling [65–67]. Table I summarizes the type IIB brane configuration. Brane tilings can be illustrated in terms of bipartite graphs on a 2-torus T^2 [68,69] and encapsulate both the field theory information and the information about the associated toric Calabi-Yau geometry. Figure 1 shows an example of a brane tiling and its associated toric Calabi-Yau 3-fold, which is in this case the cone over the zeroth Hirzebruch surface F_0 [35,38,70,71].

TABLE I. Type IIB brane configuration for brane tilings, where Σ: $P(x,y)=0$ refers to the holomorphic curve defined by the corresponding toric Calabi-Yau 3-fold and the Newton polynomial $P(x,y)$ of the associated toric diagram Δ [83,84].

	0	1	2	3	4	5	6	7	8	9
D5	✗	✗	✗	✗	·	✗	·	✗	·	·
NS5	✗	✗	✗	✗		—Σ—			·	·

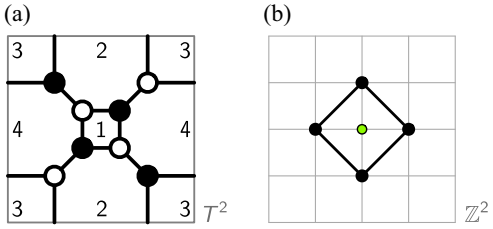

FIG. 1. (a) The brane tiling for the second phase of the zeroth Hirzebruch surface F_0, and (b) its corresponding toric diagram [35,38,70,71].

The mesonic moduli spaces [45,46,72,73] formed by the mesonic gauge invariant operators of these $4d$ $\mathcal{N}=1$ supersymmetric gauge theories with $U(1)$ gauge groups is precisely the associated toric Calabi-Yau 3-folds. When all the gauge groups of the $4d$ $\mathcal{N}=1$ supersymmetric gauge theory are $U(N)$, then the mesonic moduli space is given by the Nth symmetric product of the toric Calabi-Yau 3-fold.

The gravity dual of the $4d$ world volume theories is type IIB string theory on $\mathrm{AdS}_5 \times Y_5$, where Y_5 is the Sasaki-Einstein 5-manifold that forms the base of the associated toric Calabi-Yau 3-fold [37–39,61–64]. These $4d$ $\mathcal{N}=1$ supersymmetric gauge theories are known to flow at low energies to a superconformal fixed point. Under a procedure known as a-maximization [40–42], the superconformal R-charges of the $4d$ theory are determined. This procedure involves the maximization of the trial a-charge, which takes the form

$$a(R;Y_5)=\frac{3}{32}(3\mathrm{Tr}R^3-\mathrm{Tr}R). \tag{2.1}$$

The maximization procedure gives the value of the central charge of the superconformal field theory at the conformal fixed point.

Under the AdS/CFT correspondence [37–39], the central charge is directly related to the minimized volume of the corresponding Sasaki-Einstein 5-manifold Y_5 [43,44]. We have

$$a(R;Y_5)=\frac{\pi^3 N^2}{4V(R;Y_5)}, \tag{2.2}$$

where the R-charges R and as a result the volume function $V(R;Y_5)$ can be expressed in terms of Reeb vector components b_i of the corresponding Sasaki-Einstein 5-manifold [25,26]. We can reverse the statement saying that computing the minimum volume,

$$V_{\mathrm{min}}=\mathrm{min}_{b_i}V(b_i;Y_5), \tag{2.3}$$

is equivalent to obtaining the maximum value of the central charge $a(R;Y_5)$. This correspondence is true for all $4d$ theories living on a stack of N D3-branes probing toric

Calabi-Yau 3-folds and has been checked extensively in various examples [40–42].

In this work, we will focus on the toric Calabi-Yau 3-folds and the corresponding Sasaki-Einstein 5-manifold Y_5, with particular emphasis on the minimum volume $V_{\min}$ of the Sasaki-Einstein 5-manifolds Y_5. Building on the pioneering work of [2], this work proposes the use of more advanced machine learning techniques. In particular, we introduce machine learning regularization by using the Lasso method [60] in order to identify an explicit formula for the minimum volume $V_{\min}$ for Sasaki-Einstein 5-manifolds Y_5. We expect to be able to write the minimum volume formula in terms of features obtained from the toric diagram of the corresponding toric Calabi-Yau 3-folds. The use of machine learning regularization allows us to eliminate parameters, reducing the necessary parameters for the volume formula to a manageable amount that is interpretable, presentable and reusable.

Before discussing these machine learning techniques, let us first review in the following section the computation of the volume functions for toric Calabi-Yau 3-folds using Hilbert series.

III. HILBERT SERIES AND CALABI-YAU VOLUMES

Given $\mathcal{X}$ as a cone over a projective variety X, where X is realized as an affine variety in $\mathbb{C}$, the Hilbert series [45,46] is the generating function for the dimension of the graded pieces of the coordinate ring

$$\mathbb{C}[x_1, \ldots, x_k]/\langle f_i \rangle, \tag{3.1}$$

where f_i are the defining polynomials of X. Accordingly, the Hilbert series takes the general form

$$g(t; \mathcal{X}) = \sum_{i=0}^{\infty} \dim_{\mathbb{C}}(X_i) t^i. \tag{3.2}$$

For $4d$ $\mathcal{N} = 1$ supersymmetric gauge theories given by brane tilings [65–67], we have an associated toric Calabi-Yau 3-fold $\mathcal{X}$, which becomes the mesonic moduli space [45,46,72,73] of the $4d$ $\mathcal{N} = 1$ supersymmetric gauge theory when the gauge groups are all $U(1)$. The corresponding Hilbert series is the generating function of mesonic gauge invariant operators that form the mesonic moduli space. For the purpose of the remaining discussion, we will consider the $4d$ $\mathcal{N} = 1$ supersymmetric gauge theories given by brane tilings as Abelian theories with $U(1)$ gauge groups.

Following the forward algorithm for brane tilings [35], we can use gauged linear sigma model (GLSM) fields [72] given by perfect matchings p_α [65,66] of the brane tilings in order to express the mesonic moduli space of the Abelian $4d$ $\mathcal{N} = 1$ supersymmetric gauge theory as the following symplectic quotient:

$$\mathcal{X} = {}^{\mathrm{Irr}}\mathcal{F}^\flat // Q_D = (\mathbb{C}[p_\alpha] // Q_F) // Q_D, \tag{3.3}$$

where $\mathrm{Irr}\mathcal{F}^\flat$ is the largest irreducible component, also known as the coherent component, of the master space $\mathcal{F}^\flat$ [74–76] of the $4d$ $\mathcal{N} = 1$ supersymmetric gauge theory. The master space is the spectrum of the coordinate ring generated by the chiral fields encoded in p_α and quotiented by the F-term relations encoded in Q_F. In Eq. (3.3), Q_F is the F-term charge matrix summarizing the $U(1)$ charges originating from the F-terms, and Q_D is the D-term charge matrix which summarizes the $U(1)$ gauge charges on perfect matchings p_α.

Following the symplectic quotient description of the mesonic moduli space in Eq. (3.3), the Hilbert series can be obtained by solving the Molien integral [77],

$$g(y_\alpha; \mathcal{X}) = \prod_{i=1}^{c-2} \oint_{|z_i|=1} \frac{\mathrm{d}z_i}{2\pi i z_i} \prod_{\alpha=1}^{c} \frac{1}{1 - y_\alpha \prod_{j=1}^{c-3} z_j^{(Q_t)_{j\alpha}}}, \tag{3.4}$$

where c is the number of perfect matchings in the brane tiling and $Q_t = (Q_F, Q_D)$ is the total charge matrix.

References [25,26] showed that the same Hilbert series can be obtained directly from the toric diagram Δ of the toric Calabi-Yau 3-fold $\mathcal{X}$. Given that the toric diagram Δ is a convex lattice polygon on $\mathbb{Z}^2$ with an ideal triangulation $\mathcal{T}(\Delta)$ into unit subtriangles $\Delta_i \in \mathcal{T}(\Delta)$, the Hilbert series of the corresponding toric Calabi-Yau 3-fold $\mathcal{X}$ can be written as

$$g(t_i; \mathcal{X}) = \sum_{i=1}^{r} \prod_{j=1}^{n} \frac{1}{(1 - \mathbf{t}^{\mathbf{u}_{i,j}})}, \tag{3.5}$$

where $i = 1, \ldots, r$ is the index for the r unit triangles $\Delta_i \in \mathcal{T}(\Delta)$, and $j = 1, 2, 3$ is the index for the three boundary edges of each unit triangle Δ_i. For each boundary edge $e_j \in \Delta_i$, we have a three-dimensional outer normal vector $\mathbf{u}_{i,j}$ whose components are assigned the following product of fugacities,

$$\mathbf{t}^{\mathbf{u}_{i,j}} = \prod_{a}^{3} t_a^{\mathbf{u}_{i,j}(a)}, \tag{3.6}$$

where $\mathbf{u}_{i,j}(a)$ indicates the ath component of $\mathbf{u}_{i,j}$. We note that $\mathbf{u}_{i,j}$ is a three-dimensional vector because the defining vertices of Δ and Δ_i are all on a plane at height $z = 1$ such that their coordinates are of the form $(x, y, 1)$. As a result, the vectors $\mathbf{u}_{i,j}$ corresponding to edge $e_j \in \Delta_j$ are normal to the three-dimensional surface given by the vectors connecting the origin (0, 0, 0) to the two bounding vertices of $e_j \in \Delta_j$.

It is important to note that the fugacities t_1, t_2, t_3 in Eq. (3.6) relate to the components of normal vectors $\mathbf{u}_{i,j}$, and therefore depend on the triangulation and the particular

instance in a given $GL(2,\mathbb{Z})$ toric orbit of a toric diagram on the $z=1$ plane. In comparison, the fugacities y_α in Eq. (3.4) refer to the GLSM fields p_α given by perfect matchings of the corresponding brane tiling. Since perfect matchings can be mapped directly to chiral fields in the $4d$ $\mathcal{N}=1$ supersymmetric gauge theory, the fugacities y_α in Eq. (3.4) can be mapped to fugacities counting global symmetry charges carried by chiral fields in the $4d$ theory. Because both Hilbert series from Eq. (3.4) and Eq. (3.5) refer to the same toric Calabi-Yau 3-fold $\mathcal{X}$, there exists a fugacity map between y_α and t_1, t_2, t_3 that identifies the two Hilbert series with each other.

For the rest of the discussion, let us consider Hilbert series for toric Calabi-Yau 3-folds $\mathcal{X}$ that are in terms of fugacities t_1, t_2, t_3 corresponding to coordinates of the normal vectors $\mathbf{u}_{i,j} \in \mathbb{Z}^3$ of the toric diagram Δ. Given the Hilbert series $g(t_i;\mathcal{X})$, we can obtain the volume function [25,26] of the Sasaki-Einstein 5-manifold Y_5 using,

$$V(b_i;Y_5) = \lim_{\mu\to 0}\mu^3 g(t_i=\exp[-\mu b_i];\mathcal{X}), \tag{3.7}$$

where b_i are the Reeb vector components with $i=1,\ldots 3$. We note that the Reeb vector $\mathbf{b}=(b_1,b_2,b_3)$ is always in the interior of the toric diagram Δ and can be chosen such that one of its components is set to

$$b_3 = 3, \tag{3.8}$$

for toric Calabi-Yau 3-folds $\mathcal{X}$. We further note that the limit in Eq. (3.7) takes the leading order in μ in the expansion for $g(t_i=\exp[-\mu b_i];\mathcal{X})$, which is shown to refer to the volume of the Sasaki-Einstein base Y_5 in [25,26].

Let us consider in the following paragraph an example of the computation of the volume function in terms of Reeb vector components b_i for the Sasaki-Einstein base of the cone over the zeroth Hirzebruch surface F_0 [35,38,70,71].

A. Example: F_0

The toric diagram, its triangulation and the outer normal vectors $\mathbf{u}_{i,j}$ for the cone over the zeroth Hirzebruch surface F_0 [35,38,70,71] are shown in Fig. 2(a). The cone over the zeroth Hirzebruch surface F_0 is an interesting toric Calabi-Yau 3-fold because it has two distinct corresponding $4d$ $\mathcal{N}=1$ supersymmetric gauge theories represented by two distinct brane tilings that are related by Seiberg duality [35,78,79]. One of the brane tilings is shown in Fig. 1.

Using the outer normal vectors $\mathbf{u}_{i,j}$ for each of the four unit subtriangles Δ_i of the toric diagram for F_0 in Fig. 2(b), we can use Eq. (3.5) to write down the Hilbert series,

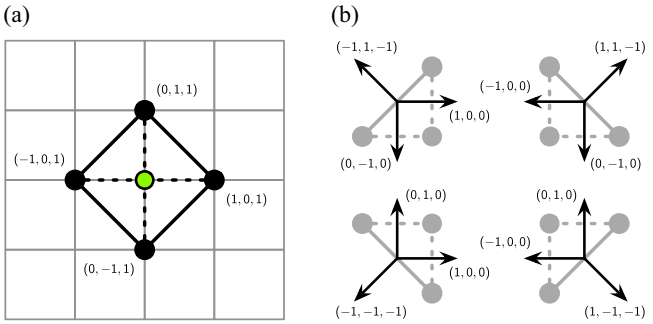

FIG. 2. (a) The triangulated toric diagram for the zeroth Hirzebruch surface F_0, and (b) the corresponding normal vectors $\mathbf{u}_{i,j}$ for each unit triangle Δ_i in the triangulation.

$$\begin{aligned} g(t_i;F_0) = &\frac{1}{(1-t_1)(1-t_2^{-1})(1-t_1^{-1}t_2t_3^{-1})} \\ &+\frac{1}{(1-t_1^{-1})(1-t_2^{-1})(1-t_1t_2t_3^{-1})} \\ &+\frac{1}{(1-t_1)(1-t_2)(1-t_1^{-1}t_2^{-1}t_3^{-1})} \\ &+\frac{1}{(1-t_1^{-1})(1-t_2)(1-t_1t_2^{-1}t_3^{-1})}. \end{aligned} \tag{3.9}$$

Using the limit in Eq. (3.7), we can derive the volume function of the Sasaki-Einstein base directly from the Hilbert series as follows:

$$\begin{aligned} &V(b_i;F_0) \\ &= \frac{24}{(b_1-b_2-3)(b_1-b_2+3)(b_1+b_2-3)(b_1+b_2+3)}, \end{aligned} \tag{3.10}$$

where $b_3=3$. When we find the global minimum of the volume function $V(b_i;F_0)$, we obtain

$$V_{\min} = \min_{b_i} V(b_i;F_0) = \frac{8}{27} \simeq 0.29630, \tag{3.11}$$

up to five decimal points, which occurs at critical Reeb vector components $b_1^* = b_2^* = 0$. In the remainder of this work, we will maintain a precision level of five decimal points for all numerical measurements.

IV. FEATURES OF TORIC DIAGRAMS AND REGRESSION

The aim of this work is to identify an expression for the minimum volume $V_{\min}$ of Sasaki-Einstein 5-manifolds Y_5 in terms of parameters that we know from the corresponding toric Calabi-Yau 3-folds $\mathcal{X}$. We refer to these parameters as features, denoted as x_a, of the toric Calabi-Yau 3-fold $\mathcal{X}$.

Assuming that we have N_x features x_a for a given toric Calabi-Yau 3-fold, the proposal in [2] states that we can write down a candidate linear function for the inverse minimum volume in terms of these features as follows:

$$1/\hat{V}_{\text{min}}(x_a^j) \equiv \hat{y}^j = \beta_0 + \sum_{a=1}^{N_x} \beta_a x_a^j, \tag{4.1}$$

where β_0 and β_a are real coefficients, and j labels the particular toric Calabi-Yau 3-fold $\mathcal{X}^j$ with its corresponding toric diagram $\Delta^j \in \mathbb{Z}^2$.

Let us refer to the inverse of the actual minimum volume obtained by volume minimization as $1/V_{\text{min}}^j \equiv y^j$ for a given toric Calabi-Yau 3-fold $\mathcal{X}^j$. If for a set S of $N = |S|$ toric Calabi-Yau 3-folds $\mathcal{X}^j$, we know the actual minimum volumes V_{min}^j via volume minimization, then we can calculate the following residual sum of squares of the difference between the inverses of the actual and the expected minimum volumes for the entire set S,

$$\begin{aligned}\mathcal{L} &= \frac{1}{2N}\sum_{j=1}^{N=|S|}(y^j - \hat{y}^j)^2 \\ &= \frac{1}{2N}\sum_{j=1}^{N}\left(1/V_{\text{min}}^j - \beta_0 - \sum_{a=1}^{N_x}\beta_a x_a^j\right)^2.\end{aligned} \tag{4.2}$$

Here, $\mathcal{L}$ can be considered as a loss function [80] that evaluates the performance of the candidate function for the minimum volume in Eq. (4.1). In multiple linear regression [48–52], as initially proposed in [2], the optimization task is to minimize the loss function in Eq. (4.2) for a given dataset S of toric Calabi-Yau 3-folds,

$$\text{argmin}_{\beta_0,\beta_a}\mathcal{L}. \tag{4.3}$$

In [2], multiple linear regression was used to obtain a candidate minimum volume function using the following feature set:

$$x_a^j \in \{f_1, f_2, f_3, f_1f_2, f_1f_3, \ldots, f_1^2, f_2^2, f_3^2\}^j, \tag{4.4}$$

where

$$f_1 = I, \qquad f_2 = E, \qquad f_3 = V, \tag{4.5}$$

corresponding respectively to the number of internal lattice points in Δ^j, the number of boundary lattice points in Δ^j, and the number of vertices that form the extremal corner points in Δ^j, for a given toric Calabi-Yau 3-fold $\mathcal{X}^j$. Under Pick's theorem [81], these features are related as follows:

$$A = I + E/2 - 1, \tag{4.6}$$

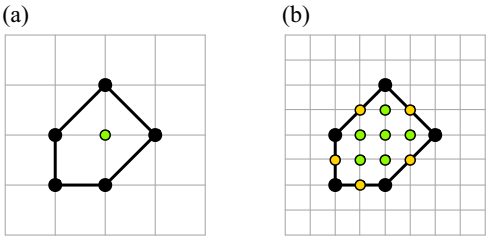

FIG. 3. (a) The toric diagram Δ_1 for the cone over dP_1, and (b) the corresponding 2-enlarged toric diagram Δ_2 with $n = 2$.

where A is the area of the toric diagram Δ, with the area of the smallest unit triangle in $\mathbb{Z}^2$ having $A = 1/2$.

With a dataset S of $N = 15,147$ toric Calabi-Yau 3-folds, the work in [2] showed that the candidate linear function in Eq. (4.1) with features given by Eq. (4.4) is able to estimate the inverse minimum volume with an expected percentage relative error of 2.2%. In this work, we expand upon the accomplishments of [2] by introducing novel features that describe toric Calabi-Yau 3-folds, augmenting the datasets for toric Calabi-Yau 3-folds, and applying machine learning techniques incorporating regularization. These improvements are designed to address some of the shortcomings of the work in [2] as well as give explicit interpretable formulas for the minimum volume for toric Calabi-Yau 3-folds.

A. New features

We introduce several new features that describe a toric Calabi-Yau 3-fold and are obtained from the corresponding toric diagram Δ. By defining the n-enlarged toric diagram as

$$\Delta_n = \{nv = (nx, ny) | v = (x, y) \in \Delta\}, \tag{4.7}$$

where $n \in \mathbb{Z}^+$ and $v = (x, y) \in \mathbb{Z}^2$ are the coordinates of the vertices in the original toric diagram Δ. We note that $\Delta_1 = \Delta$ and Fig. 3 shows as an example the 2-enlarged toric diagram Δ_2 for the cone over dP_1 These n-enlarged toric diagrams Δ_n also appeared in [82] for the study of Hodge numbers of Calabi-Yau manifolds that are constructed as hypersurfaces in toric varieties given by Δ.

Using the n-enlarged toric diagram Δ_n^j for a given toric Calabi-Yau 3-fold $\mathcal{X}^j$, we can now refer to the area of Δ_n as A_n, the number of internal lattice points of Δ_n as I_n, and the number of boundary lattice points in Δ_n as E_n. We further note that the number of vertices V_n corresponding to extremal corner points in Δ_n is the same for V in Δ for all n, i.e., $V_n = V$.

In our work, we use features of a toric Calabi-Yau 3-fold $\mathcal{X}^j$ that are composed from members of the following set:

$$\{A, V, E, I_n\}^j, \tag{4.8}$$

TABLE II. For training the machine learning models, we make use of 4 sets S_m of toric diagrams with different sizes $|S_m|$.

Set	Description	$\lvert S_m\rvert$
S_{1a}	All polytopes 5×5 lattice box	15,327
S_{1b}	All polytopes $r=3.5$ circle	31,324
S_{2a}	Selected polytopes 30×30 lattice box	202,015
S_{2b}	Selected polytopes $r=15$ circle	201,895

where $n=1,\ldots,7$. These are defined through the corresponding toric diagram Δ^j and its corresponding n-enlarged toric diagram Δ^j_n. Through the application of machine learning regularization, our objective is to differentiate between features that contribute to the expression for the minimum volume associated with a toric Calabi-Yau 3-fold and those that do not.

B. New sets of toric Calabi-Yau 3-folds

The aim of this work is to make use of machine learning with regularization in order to identify an interpretable formula that accurately estimates the minimum volume of Sasaki-Einstein 5-manifolds corresponding to toric Calabi-Yau 3-folds. The interpretability of the minimum volume formula is achieved by the lowest possible number of features on which the formula depends on. In order to train such a regularized machine learning model, we establish four sets S_m of toric Calabi-Yau 3-folds $\mathcal{X}^j$, for which the corresponding minimum volumes are known. These sets S_m are summarized in Table II and are defined as follows:

(i) S_{1a}: This set consists of toric Calabi-Yau 3-folds whose toric diagrams fit into a 5×5 lattice box in $\mathbb{Z}^2$ as illustrated in Fig. 4(a). This set contains a certain degree of redundancy given that convex lattice polygons related by a $GL(2,\mathbb{Z})$ transformation on their vertices refer to the same toric Calabi-Yau 3-fold. Accordingly, we restrict ourselves to toric diagrams Δ^j that give unique combinations of the form $(1/V^j_{\min}, V^j, E^j, I^j)$. This results in a dataset of $|S_{1a}|=15,327$ distinct toric diagrams with unique inverse minimum volumes $1/V^j_{\min}$ up to six decimal points.

(ii) S_{1b}: The second set consists of toric Calabi-Yau 3-folds whose toric diagrams fit inside a circle centered at the origin (0, 0) on the $\mathbb{Z}^2$ lattice with radius $r=3.5$ as illustrated in Fig. 4(b). By imposing the condition that we want $GL(2,\mathbb{Z})$-distinct toric diagrams Δ^j with unique combinations of the form $(1/V^j_{\min}, V^j, E^j, I^j)$, we obtain $|S_{1b}|=31,324$ toric diagram for this set.

(iii) S_{2a}: For this set, we choose randomly 300,000 toric diagrams that fit into a 30×30 lattice box in $\mathbb{Z}^2$. By imposing the condition that the toric diagrams Δ^j have unique combinations of the form $(1/V^j_{\min}, V^j, E^j, I^j)$, we obtain $|S_{2a}|=202,015$ toric diagram for this set.

(iv) S_{2b}: For this set, we choose randomly 300,000 toric diagrams that fit into a circle centered at the origin (0, 0) on the $\mathbb{Z}^2$ lattice with radius $r=15$. By imposing the condition that the toric diagrams Δ^j have unique combinations of the form $(1/V^j_{\min}, V^j, E^j, I^j)$, we obtain $|S_{2b}|=201,895$ toric diagram for this set.

The distribution of inverse minimum volumes $1/V_{\min}$ for the above sets of toric diagrams is illustrated together with the mean inverse minimum volume $\bar{y}=\langle 1/V_{\min}\rangle=\frac{1}{|S_m|}\sum_{j=1}^{|S_m|}1/V^j_{\min}$ in Fig. 5. In the following sections, we make use of regularized machine learning in order to identify functions that optimally estimate the inverse minimum volume $1/V_{\min}$ in each of the above datasets.

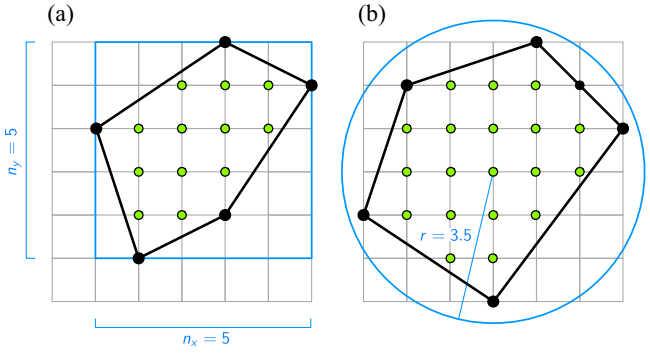

FIG. 4. (a) Toric diagrams in datasets S_{1a} and S_{2a} are constrained by a $n_x\times n_y$ lattice box in $\mathbb{Z}^2$, whereas (b) toric diagrams in datasets S_{1b} and S_{2b} are constrained by a circle of radius r with the center at $(0,0)\in\mathbb{Z}^2$.

C. Feature analysis with principal component analysis

If we restrict ourselves to features of the form $\{(f^j_u)^a(f^j_v)^b\,|1\le a+b\le 2, a,b\in\mathbb{Z}^+\}$ with $f^j_u\in\{A,V,E,I_n\}^j$, where $n=1,\ldots,7$, we obtain a collection of 65 dimensional feature vectors for each of the datasets S_m. The relative statistical relevance of these features can be measured by obtaining the eigenvector components of the covariance matrix using a principal component analysis (PCA) of the feature vectors for each dataset S_m. Focusing on the first principal component, which captures a relative variance of approximately 9% for each of the datasets, the square components of the corresponding eigenvector measure the relative relevance of the associated feature towards this principal component. Table III shows these values for the features of the form $(f^j_u)^a(f^j_v)^b$ according to the four datasets S_m. We see from this analysis that feature combinations involving V and E have statistically the lowest relevance. In the following work, we propose the use of

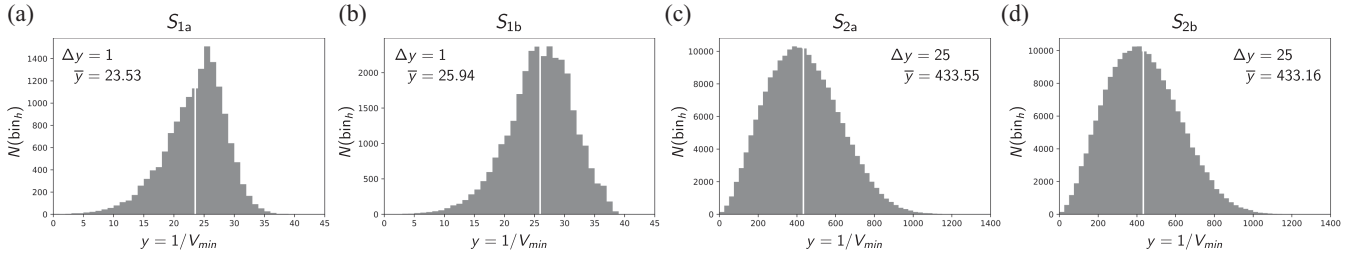

FIG. 5. The distribution of expected minimum volumes $y = 1/V_{\text{min}}$ for the datasets (a) S_{1a}, (b) S_{1b}, (c) S_{2a}, and (c) S_{2b}. The mean expected value $\bar{y}$ is indicated by a white line. The histograms for values of $y = 1/V_{\text{min}}$ are obtained for bin sizes Δy with the number of toric diagrams in bin_h given by $N(\text{bin}_h)$.

machine learning regularization as an alternative scheme of feature selection for the minimum volume of toric Calabi-Yau 3-folds.

D. Machine learning models and regularization

In order to obtain a function for the minimum volume of Sasaki-Einstein 5-manifolds corresponding to toric Calabi-Yau 3-folds in terms of features obtained from the corresponding toric diagrams, we make use of the following machine learning models:

(i) *Polynomial regression (PR).* We make use of polynomial regression [85], where the relationship between the feature variables x_a^j and the predicted variable $\hat{y}^j$, is given by

$$\hat{y}^j = \beta_0 + \sum_{a=1}^{N_x} \beta_a x_a^j, \tag{4.9}$$

where β_0 and β_a are real coefficients, N_x is the number of features, and j labels the particular sample in the data set that is used to train this machine learning model. In our case, the data set consists of toric Calabi-Yau 3-folds $\mathcal{X}^j$, where the corresponding minimum volume V_{min}^j is given by $y = 1/V_{\text{min}}^j$. Here we note that the features x_a^j are taken from the set $\{(f_u^j)^a (f_v^j)^b | 1 \leq a + b \leq 2, a, b \in \mathbb{Z}^+\}$ with $f_u^j \in \{A, V, E, I_n\}^j$, where $n = 1, \ldots, 7$.

(ii) *Logarithmic regression (LR).* We make use of logarithmic regression [85] in order to help linearize relationships between features x_a^j that are potentially multiplicative in their contribution towards the predicted variable $\hat{y}^j$. To be more precise, we make use of a log-log model where we log-transform both the predicted variable $\hat{y}^j$ and the features x_a^j. The predicted variable is then given by

$$\log(\hat{y}^j) = \beta_0 + \sum_{a=1}^{N_x} \beta_a \log(x_a^j), \tag{4.10}$$

TABLE III. Principal component analysis of feature vectors with components of the form $(f_u^j)^a (f_v^j)^b$ with $1 \leq a + b \leq 2$, $a, b \in \mathbb{Z}^+$ and $f_u^j \in \{A, V, E, I_n\}^j$, where $n = 1, \ldots, 7$. For each feature, we show the square value of the corresponding component of the first eigenvector of the PCA covariance matrix, which measures relative relevance (in %) of the feature with respect to the first principal component. The first principal component has a relative variance of approximately 9% for each of the datasets S_{1a}, S_{1b}, S_{2a} and S_{2b}.

Dataset	A	V	E	I	I_2	I_3	I_4	I_5	I_6	I_7	A^2	AV	AE	AI	AI_2	AI_3	AI_4	AI_5	AI_6	AI_7	V^2	VE
S_{1a}	1.685	0.511	0.661	1.505	1.638	1.662	1.671	1.675	1.677	1.679	1.710	1.448	1.348	1.664	1.702	1.707	1.709	1.709	1.709	1.710	0.489	0.862
S_{1b}	1.672	0.789	0.362	1.549	1.636	1.654	1.660	1.664	1.666	1.667	1.701	1.542	1.302	1.655	1.687	1.693	1.696	1.697	1.698	1.699	0.782	0.746
S_{2a}	1.730	0.422	0.167	1.716	1.724	1.726	1.727	1.728	1.728	1.729	1.795	1.637	0.884	1.787	1.791	1.792	1.793	1.793	1.794	1.794	0.420	0.324
S_{2b}	1.731	0.430	0.163	1.717	1.724	1.727	1.728	1.728	1.729	1.729	1.794	1.637	0.886	1.787	1.791	1.792	1.792	1.793	1.793	1.793	0.427	0.320

Dataset	VI	VI_2	VI_3	VI_4	VI_5	VI_6	VI_7	E^2	EI	EI_2	EI_3	EI_4	EI_5	EI_6	EI_7	I^2	II_2	II_3	II_4	II_5	II_6	II_7
S_{1a}	1.417	1.456	1.457	1.456	1.455	1.455	1.454	0.638	1.566	1.456	1.419	1.401	1.390	1.383	1.378	1.534	1.616	1.635	1.644	1.648	1.651	1.653
S_{1b}	1.564	1.566	1.560	1.557	1.554	1.552	1.551	0.333	1.541	1.427	1.386	1.365	1.352	1.344	1.338	1.568	1.620	1.633	1.639	1.642	1.645	1.646
S_{2a}	1.633	1.636	1.636	1.636	1.636	1.636	1.636	0.161	0.920	0.902	0.896	0.893	0.891	0.890	0.889	1.779	1.783	1.785	1.785	1.786	1.786	1.786
S_{2b}	1.634	1.636	1.636	1.637	1.637	1.637	1.637	0.155	0.922	0.904	0.898	0.895	0.894	0.892	0.891	1.779	1.783	1.784	1.785	1.785	1.786	1.786

Dataset	I_2^2	I_2I_3	I_2I_4	I_2I_5	I_2I_6	I_2I_7	I_3^2	I_3I_4	I_3I_5	I_3I_6	I_3I_7	I_4^2	I_4I_5	I_4I_6	I_4I_7	I_5^2	I_5I_6	I_5I_7	I_6^2	I_6I_7	I_7^2
S_{1a}	1.675	1.687	1.692	1.694	1.696	1.697	1.696	1.700	1.702	1.703	1.704	1.703	1.704	1.705	1.706	1.706	1.707	1.707	1.707	1.708	1.708
S_{1b}	1.662	1.672	1.676	1.679	1.680	1.681	1.681	1.685	1.687	1.688	1.689	1.688	1.690	1.691	1.692	1.692	1.693	1.694	1.694	1.695	1.695
S_{2a}	1.788	1.789	1.789	1.790	1.790	1.790	1.790	1.791	1.791	1.791	1.791	1.791	1.792	1.792	1.792	1.792	1.792	1.792	1.792	1.793	1.793
S_{2b}	1.787	1.788	1.789	1.789	1.789	1.790	1.789	1.790	1.790	1.791	1.791	1.791	1.791	1.791	1.791	1.791	1.792	1.792	1.792	1.792	1.792

where β_0 and β_a are real coefficients, and N_x is the number of log-transformed features of the form $\log(x_a^j)$. The label j corresponds to a particular toric Calabi-Yau 3-fold $\mathcal{X}^j$ whose corresponding minimum volume $V_{\min}^j$ is given by $y^j = 1/V_{\min}^j$. Here we note that the log-transformed features of the form $\log(x_a^j)$ are taken from the set $\{(\log(f_u^j))^a(\log(f_v^j))^b | 1 \le a+b \le 2, a,b \in \mathbb{Z}^+\}$ with $f_u^j \in \{A, V, E, I_n\}^j$, where $n = 3, \ldots, 7$. Here, we do not make use of I_1 and I_2.

When we introduce regularization [58,59] into polynomial regression and logarithmic regression, we minimize the following loss function between the predicted variable $\hat{y}^j$ and the expected variable y,

$$\mathcal{L} = \frac{1}{2N}\sum_{j=1}^{N}(y^j - \hat{y}^j)^2 + \Delta\mathcal{L}, \tag{4.11}$$

where $\Delta\mathcal{L}$ is the regularization term in the loss function. The loss function in Eq. (4.11) is iteratively minimized during the optimization process and we set for all following computations the maximum number of iterative steps to be $N_{\max} = 10,000$. The precise form of the regularization term in the loss function as well as the different regularization schemes in machine learning are discussed in the following section.

V. LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO) AND REGULARIZATION

Lasso [60] is a machine learning regularization technique primarily employed to prevent overfitting in supervised machine learning. However, it can also be utilized for feature selection. In our work, the overarching goal in employing Lasso is to introduce a machine learning model capable of delivering optimal predictions for the minimum volume for toric Calabi-Yau 3-folds while using the fewest features from the training dataset. For problems such as the one considered in this work, it is quintessential to be able to obtain formulas with a small number of parameters. As a result, using Lasso is particularly suited for discovering new mathematical formulas such as the one aimed for in this work for the minimum volume for toric Calabi-Yau 3-folds.

In the following section, we give a brief overview of several regularization schemes including Lasso in the context of supervised machine learning for the minimum volume formula for toric Calabi-Yau 3-folds.

A. Regularization

Regularization in machine learning is a technique usually used to avoid overfitting the dataset during model training. This is done by adding a penalty term in the loss function. The introduction of the added regularization term $\Delta\mathcal{L}$, resulting in an updated loss function of the form,

$$\mathcal{L} + \Delta\mathcal{L}, \tag{5.1}$$

serves the purpose of constraining the possible parameter values within the supervised machine learning model. In the case of multiple linear regression as first introduced in [2] and reviewed in Sec. IV, these parameters would be the real coefficients β_0 and β_a in the candidate linear function in Eq. (4.1) for the expected minimum volume given by $\hat{y}^j = 1/\hat{V}_{\min}^j$. By restricting the values for these parameters, regularization effectively makes it harder for the supervised machine learning model to give a candidate function for the minimum volume $V_{\min}$ with many terms in the function. This prevents the machine learning model to overfit the dataset of minimized volumes for toric Calabi-Yau 3-folds.

Let us review the following three regularization schemes:

(i) *L1 Regularization (Lasso).* This regularization scheme [60] adds the following linear regularization term to the loss function of the regression model,

$$\Delta\mathcal{L}_{\mathrm{L1}} = \alpha\sum_{a=1}^{N_x}|\beta_a|, \tag{5.2}$$

where β_a are the real parameters of the regression model. α is a real regularization parameter. Increasing the value of α has the effect of increasing the strength of the L1 regularization.

(ii) *L2 Regularization (Ridge).* Another regularization scheme is known as Ridge regularization or L2 regularization [86]. It adds the following quadratic regularization term to the loss function of the regression model,

$$\Delta\mathcal{L}_{\mathrm{L2}} = \alpha\sum_{a=1}^{N_x}\beta_a^2, \tag{5.3}$$

where β_a are the real parameters of the regression model and α is again the real regularization parameter.

(iii) *Elastic net (L1 and L2).* Elastic net [87] is a combination of L1 (Lasso) and L2 (Ridge) regularization and adds the following regularization terms to the loss function,

$$\Delta\mathcal{L}_{\mathrm{L1,L2}} = \alpha_1\sum_{a=1}^{N_x}|\beta_a| + \alpha_2\sum_{a=1}^{N_x}\beta_a^2, \tag{5.4}$$

where α_1 and α_2 are relative real regularization parameters that regulate the proportion of L1 regularization and L2 regularization in this regularization scheme.

Amongst these regularization schemes in supervised machine learning, we are going to mainly focus on Lasso and L1 regularization for the remainder of this work. While all three regularization schemes share the common goal of constraining the range of values for the model parameters β_a, it is noteworthy that only Lasso possesses the unique property of inducing sparsity among the model parameters, resulting in the complete elimination of certain parameters during the training process.

There are several arguments why Lasso enables the complete elimination of some of the model parameters and the corresponding features in the candidate function for the minimum volume $V_{\min}$ for toric Calabi-Yau 3-folds. In order to illustrate this, let us consider the case with $N_x = 2$ features x_1^j and x_2^j, for which the L1 and L2 regularization terms take respectively the following form,

$$\Delta\mathcal{L}_{\mathrm{L1}} = \alpha(|\beta_1| + |\beta_2|), \qquad \Delta\mathcal{L}_{\mathrm{L2}} = \alpha(\beta_1^2 + \beta_2^2). \tag{5.5}$$

If we assume that under optimization, the regularization terms reach a value $\Delta\mathcal{L}_{\mathrm{L1}} = \epsilon$ and $\Delta\mathcal{L}_{\mathrm{L2}} = \epsilon$ for $\alpha > 0$ and $\epsilon \in \mathbb{R}$, we can draw the parametric plots for the two regularization terms as shown in Fig. 6 [58]. We can see from the plots in Fig. 6 that for L1 regularization, the minimum of the total loss function is more likely achieved when one of the two parameters β_1 or β_2 approaches 0. This is in part due to the absolute values taken for the parameters in the linear L1 regularization term.

As a result, Lasso regularization is particularly suited for feature selection and parameter elimination in regression models. In our work, we employ L1 regularization to derive a formula for the minimum volume $V_{\min}$ of Sasaki-Einstein 5-manifolds corresponding to toric Calabi-Yau 3-folds that is interpretable, presentable and reusable.

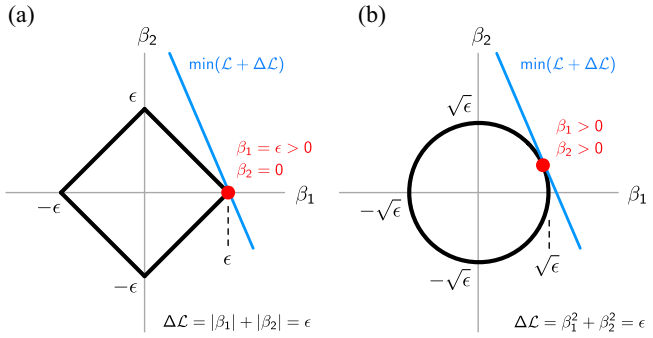

FIG. 6. Parametric plots for β_1 and β_2 for a 2-parameter model [58]. (a) In L1 regularization (Lasso), the minimum of the regularized loss function $\min(\mathcal{L} + \Delta\mathcal{L})$ is more likely to be located when one of the parameters vanishes, in comparison to the case (b) in L2 regularization (Ridge) where the minimum of the regularized loss function $\min(\mathcal{L} + \Delta\mathcal{L})$ is equally more likely located at nonzero values of the parameters. This illustrates that L1 regularization is more suited in eliminating parameters under optimization.

VI. CANDIDATES FOR MINIMUM VOLUME FUNCTIONS

In this work, our aim is to apply Lasso regularization in order to identify explicit formulas for the minimum volume for toric Calabi-Yau 3-folds. By doing so, our aim is to maximize the accuracy of the formulas that we find while minimizing the number of parameters the formulas depend on, making them interpretable and readily presentable.

A. Parameter sparsity vs accuracy

Like in all regression problems, we introduce as a measure of how well the model fits the observed data using the R^2-score [58,85] given by

$$R^2 = 1 - \frac{S_{\mathrm{res}}}{S_{\mathrm{tot}}}, \tag{6.1}$$

where the residual sum of squares S_{res} is given by

$$S_{\mathrm{res}} = \sum_{j=1}^{N} (y^j - \hat{y}^j)^2 \tag{6.2}$$

and the total sum of squares S_{tot} is given by

$$S_{\mathrm{tot}} = \sum_{j=1}^{N} (y^j - \bar{y})^2. \tag{6.3}$$

Here, $\hat{y}^j$ denotes the predicted value for the minimum volume $V_{\min}^j$ given by $y^j = 1/V_{\min}^j$, whereas $\bar{y}$ denotes the mean of the expected values y^j.

We recall that the optimization problem for the L1-regularized regression model is to minimize the loss function $\mathcal{L} + \Delta\mathcal{L}_{\mathrm{L1}}$ with the L1 regularization term. As we discussed in the sections above, this optimization problem focuses on minimizing the mean squared error with a penalty for nonzero coefficients $\beta_a(\alpha)$, which depends on the regularization parameter α.

Here, we note that there is an additional optimization problem regarding the maximization of the R^2-score in Eq. (6.1) and the minimization of the number $N_{\beta_a(\alpha)}$ of nonzero coefficients $\beta_a(\alpha)$. We can formulate this additional optimization problem as follows:

$$\max_{\alpha} \left\{ R^2(\alpha) - \lambda \frac{N_{\beta_a(\alpha)}}{N_x} \right\}, \tag{6.4}$$

where $0 < N_{\beta_a(\alpha)} \leq N_x$, and the values of the coefficients $\beta_a(\alpha)$ and the $R^2(\alpha)$-score all depend on the regularization parameter α. λ is a positive hyperparameter that regulates how much we value sparsity of feature coefficients $\beta_a(\alpha)$ over the accuracy of the estimate given by $R^2(\alpha)$.

B. Candidate formulas

The candidate formulas for the minimum volume for toric Calabi-Yau 3-folds are identified by an optimal regularization parameter α^* that maximizes the R^2-score of the candidate formula and minimizes the number of nonzero coefficients $N_{\beta_a(\alpha)}$ corresponding to features in the chosen regression model. In order to identify the optimal regularization parameter α^* for the optimization problem in Eq. (6.4), we search for α^* in a given fixed range for α as specified in Figs. 7 and 8. We do the search for the optimal regularization parameter α^* for all four datasets in Table II for both L1-regularized polynomial regression and L1-regularized logarithmic regression as discussed in Secs. IV and V. The chosen L1-regularized regression models are trained for a particular value of the regularization parameter α under a fixed randomly chosen 80% training and 20% testing data split, where the corresponding R^2-score depending on α is obtained from the testing data.

Figure 7 shows respectively for datasets S_{1a} and S_{2a} plots for the L1 regularization parameter α for polynomial regression against standardized coefficients $\bar{\beta}_a(\alpha)$, against the number of nonzero coefficients $N_{\beta_a(\alpha)}$, and against the R^2-score. Here, the standardized coefficients $\bar{\beta}_a(\alpha)$ are obtained when the training is conducted over normalized features $\bar{x}_a$. When the training is completed for a specific value of α, the candidate formula for the minimum volume given by $y = 1/V_{\min}$ is obtained by reversing the normalization on the features, giving us the coefficients $\beta_a(\alpha)$ of the candidate formula. We also have Fig. 8 which shows respectively for datasets S_{1a} and S_{2a} plots for the L1 regularization parameter α for logarithmic regression against the standardized coefficients $\bar{\beta}_a(\alpha)$, the number of nonzero coefficients $N_{\beta_a(\alpha)}$ and the R^2-score. Similar plots can also be obtained for datasets S_{1b} and S_{2b} for both L1-regularized polynomial regression and L1-regularized logarithmic regression.

Overall, the plots illustrate that the identified optimal regularization parameters α^* minimize the number of nonzero coefficients $N_{\beta_a(\alpha)}$ in the formula estimating the minimum volume given by $y = 1/V_{\min}$, as well as maximize the accuracy of the formulas measured by the R^2-score. Tables IV and V summarize respectively the most optimal candidate formulas for the minimum volume given by $y = 1/V_{\min}$ under L1-regularized polynomial regression and L1-regularized logarithmic regression for the four datasets in Table II, with the corresponding optimal regularization parameters α^*, the corresponding number of nonzero coefficients $N_{\beta_a(\alpha)}$ and the R^2-score.

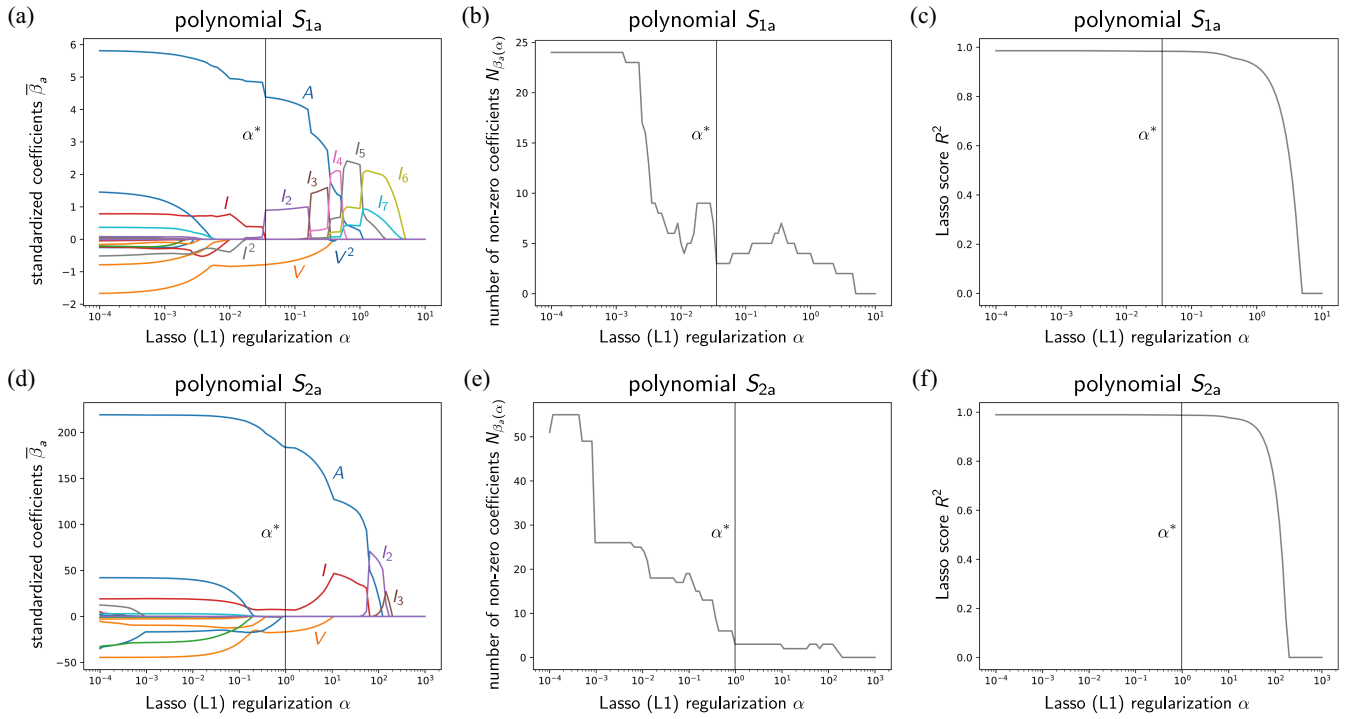

FIG. 7. The L1 (Lasso) regularization parameter α for polynomial regression on dataset S_{1a} (15,327 toric diagrams in 5×5 lattice box) against (a) the standardized coefficients $\bar{\beta}_a(\alpha)$, (b) the number of nonzero coefficients $N_{\beta_a(\alpha)}$, and (c) the corresponding $R^2(\alpha)$-score. The optimal regularization parameter α^* was found in the range $\alpha = 10^{-4}, \ldots, 10^1$ by taking steps of $\Delta\alpha \simeq 1.12202$. We also have the L1 (Lasso) regularization parameter α for polynomial regression on dataset S_{2a} (202,015 random toric diagrams in 30×30 lattice box) against (c) the standardized coefficients $\bar{\beta}_a(\alpha)$, (d) the number of nonzero coefficients $N_{\beta_a(\alpha)}$, and (e) the corresponding $R^2(\alpha)$-score. The optimal regularization parameter α^* was found in the range $\alpha = 10^{-4}, \ldots, 10^3$ by taking steps of $\Delta\alpha \simeq 1.17490$.

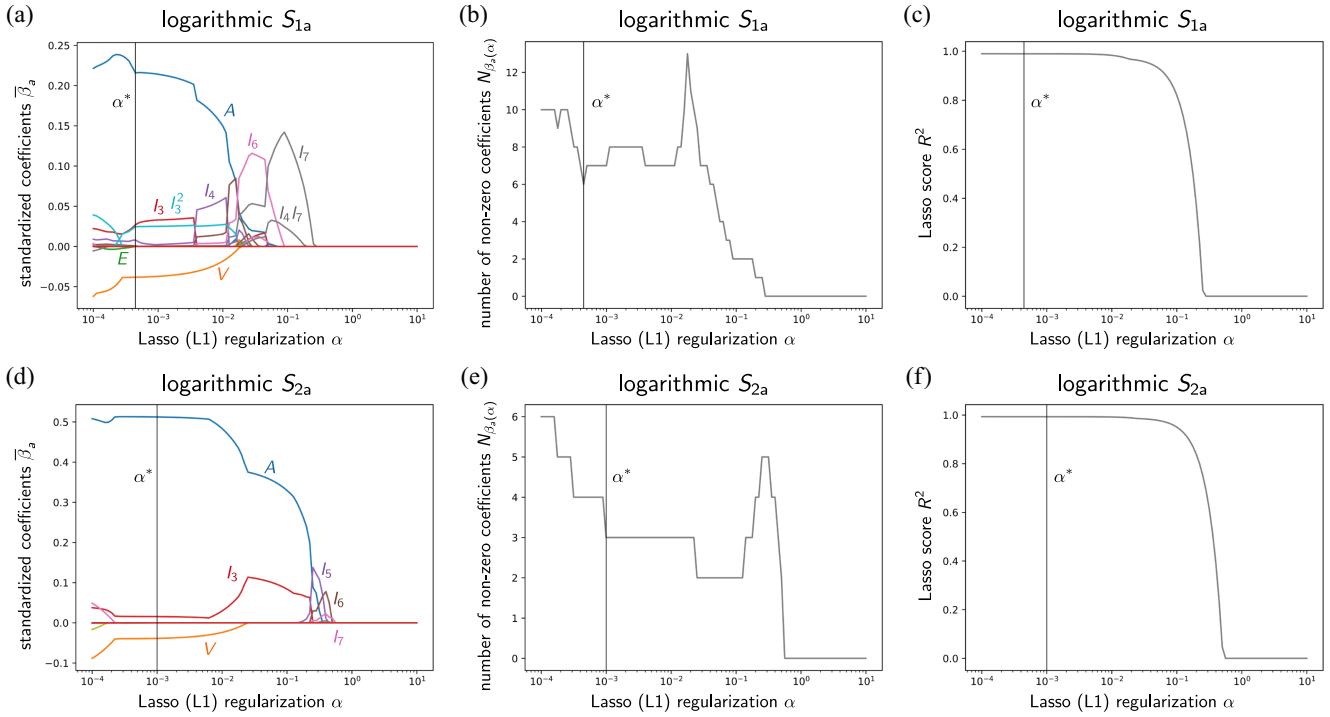

FIG. 8. The L1 (Lasso) regularization parameter α for logarithmic regression on dataset S_{1a} (15,327 toric diagrams in 5×5 lattice box) against (a) the standardized coefficients $\bar{\beta}_a(\alpha)$, (b) the number of nonzero coefficients $N_{\beta_a(\alpha)}$, and (c) the corresponding $R^2(\alpha)$-score. The optimal regularization parameter α^* was found in the range $\alpha = 10^{-4}, \ldots, 10^1$ by taking steps of $\Delta\alpha \simeq 1.12202$. We also have the L1 (Lasso) regularization parameter α for logarithmic regression on dataset S_{2a} (202,015 random toric diagrams in 30×30 lattice box) against (c) the standardized coefficients $\bar{\beta}_a(\alpha)$, (d) the number of nonzero coefficients $N_{\beta_a(\alpha)}$, and (e) the corresponding $R^2(\alpha)$-score. The optimal regularization parameter α^* was found in the range $\alpha = 10^{-4}, \ldots, 10^3$ by taking steps of $\Delta\alpha \simeq 1.17490$.

A closer look reveals that for all models, the identified optimal regularization parameters α^* results in formulas that approximate the minimum volume $y = 1/V_{\text{min}}$ extremely well for all the datasets S_{1a}, S_{1b}, S_{2a}, and S_{2b}. Overall, the L1-regularized logarithmic regression models seem to give more accurate results than the L1-regularized polynomial regression models with $N_{\beta_a(\alpha)} \leq 6$ over all datasets. In particular, L1-regularized logarithmic regression models trained on datasets S_{2a} and S_{2b} have R^2-scores above 0.99, which is exceptionally high.

Having a closer look at explicit examples of toric Calabi-Yau 3-folds in the datasets reveals however that the performances of the regularized regression models can vary between different toric Calabi-Yau 3-folds. For example, focusing on the L1-regularized logarithmic regression models trained on S_{1a} and S_{1b}, we observe that the minimum volumes given by $1/\hat{y}_{1a}^{\text{LR}}$ and $1/\hat{y}_{1b}^{\text{LR}}$ in Table V perform differently for toric diagrams with smaller areas A compared to toric diagrams with larger areas A as illustrated in Fig. 9. Similar observations can be made for the L1-regularized

TABLE IV. Optimal candidate formulas for the minimum volume for toric Calabi-Yau 3-folds given by $y = 1/V_{\text{min}}$ and obtained under L1 (Lasso) regularized polynomial regression (PR) on datasets S_{1a}, S_{1b}, S_{2a}, and S_{2b}. For each optimal candidate formula, we give the optimal regularization parameter α^* that maximizes the corresponding R^2-score and minimizes the number of nonzero coefficients N_{β_a} in the formula.

Dataset	$y = 1/V_{\text{min}}$	α^*	$N_{\beta_a(\alpha^*)}$	$R^2(\alpha^*)$
S_{1a}	$\hat{y}_{1a}^{\text{PR}} = 1.28837A - 0.71753V + 0.07208I_2 + 5.18969$	0.03548	3	0.98354
S_{1b}	$\hat{y}_{1b}^{\text{PR}} = 1.36089A - 0.61041V + 0.15561I + 5.31028$	0.01995	3	0.98697
S_{2a}	$\hat{y}_{2a}^{\text{PR}} = 1.61574A - 19.35740V + 0.06419I + 101.58972$	0.97724	3	0.98743
S_{2b}	$\hat{y}_{2b}^{\text{PR}} = 1.61494A - 19.42096V + 0.06494I + 101.84952$	0.97724	3	0.98740

TABLE V. Optimal candidate formulas for the minimum volume for toric Calabi-Yau 3-folds given by $y = 1/V_{\text{min}}$ and obtained under L1 (Lasso) regularized logarithmic regression on datasets S_{1a}, S_{1b}, S_{2a} and S_{2b}. For each optimal candidate formula, we give the optimal regularization parameter α^* that maximizes the corresponding R^2-score and minimizes the number of nonzero coefficients N_{β_a} in the formula.

Dataset	$y = 1/V_{\text{min}}$	α^*	$N_{\beta_a(\alpha^*)}$	$R^2(\alpha^*)$
S_{1a}	$\hat{y}^{\text{LR}}_{1a} = 1.97348A^{0.77011}V^{-0.21355}I_3^{0.08796}I_4^{0.02722}I_5^{0.00202}e^{0.00923(\log I_3)^2}$	0.00045	6	0.98932
S_{1b}	$\hat{y}^{\text{LR}}_{1b} = 1.75668A^{0.74154}V^{-0.182009}E^{0.00050}I_3^{0.16451}I_4^{0.00679}e^{0.00447(\log I_3)^2}$	0.00032	6	0.98992
S_{2a}	$\hat{y}^{\text{LR}}_{2a} = 2.50772A^{0.95411}V^{-0.21992}I_3^{0.02867}$	0.00112	3	0.99281
S_{2b}	$\hat{y}^{\text{LR}}_{2b} = 2.51288A^{0.95322}V^{-0.21970}I_3^{0.02898}$	0.00112	3	0.99297

logarithmic regression models trained on S_{2a} and S_{2b} as well as the L1-regularized polynomial regression models.

In summary, we can calculate the expected relative percentage errors $E[\epsilon]$ of the predicted minimum volumes given by $1/\hat{y}$ and the corresponding standard deviations $\sigma[\epsilon]$ for the L1-regularized logarithmic regression models as follows:

$$\begin{aligned} E[\epsilon^{\text{LR}}_{1a}] &= 2.158\%, & \sigma[\epsilon^{\text{LR}}_{1a}] &= 1.696\%, \\ E[\epsilon^{\text{LR}}_{1b}] &= 1.884\%, & \sigma[\epsilon^{\text{LR}}_{1b}] &= 1.545\%, \\ E[\epsilon^{\text{LR}}_{2a}] &= 3.577\%, & \sigma[\epsilon^{\text{LR}}_{2a}] &= 2.396\%, \\ E[\epsilon^{\text{LR}}_{2b}] &= 3.579\%, & \sigma[\epsilon^{\text{LR}}_{2b}] &= 2.399\%. \end{aligned} \tag{6.5}$$

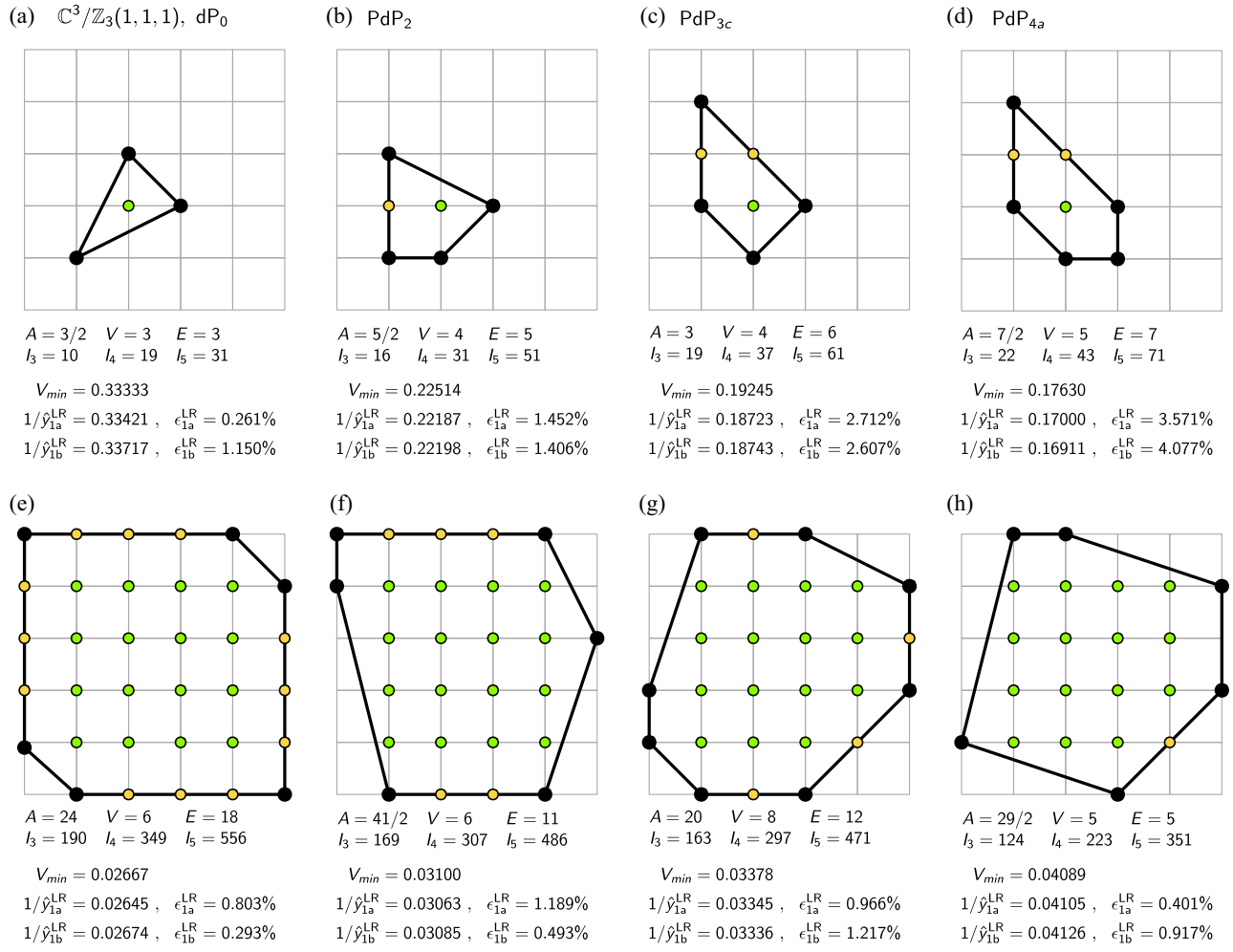

FIG. 9. The L1-regularized logarithmic regression models trained on datasets S_{1a} and S_{1b} perform better on toric diagrams with larger areas A [see selection in (e)–(h)] than for toric diagrams with smaller areas A [see selection in (a)–(d)]. The performance is measure by the relative percentage error $\epsilon(1/\hat{y})$ of the predicted minimum volume given by $1/\hat{y}$. The R^2-scores for the L1-regularized logarithmic regression models trained on datasets S_{1a} and S_{1b} are $R^2(y^{\text{LR}}_{1a}) = 0.98932$ and $R^2(y^{\text{LR}}_{1b}) = 0.98992$, respectively.

We note that the models trained on S_{2a} and S_{2b} have a larger expected relative percentage error than the ones trained on S_{1a} and S_{1b}. This is partly due to the fact that S_{2a} and S_{2b} contain randomly selected toric diagrams in a 30×30 lattice box in $\mathbb{Z}^2$ and $r = 15$ circle, respectively, whereas S_{1a} and S_{1b} contain the full set of toric diagrams in a 5×5 lattice box in $\mathbb{Z}^2$ and $r = 3.5$ circle, respectively, as defined in Table II.

We also note that the R^2-scores of the L1-regularized logarithmic regression models in Table V,

$$R^2(y_{1a}^{LR}) = 0.98932, \qquad R^2(y_{1b}^{LR}) = 0.98992,$$
$$R^2(y_{2a}^{LR}) = 0.99281, \qquad R^2(y_{2b}^{LR}) = 0.99297, \tag{6.6}$$

are overall very high and close to 1. Compared to the expected relative percentage errors in Eq. (6.5), which measure how far off predictions of the minimum volume given by $1/\hat{y}$ are, the R^2-score is a measure of the accuracy of the trained regression model. It quantifies the proportion of the variation in $y = 1/V_{\text{min}}$ that can be predicted using the features selected from the corresponding toric diagrams of the toric Calabi-Yau 3-folds.

VII. DISCUSSIONS AND CONCLUSIONS

With this work, we demonstrated that employing regularization in machine learning models can effectively address the limitations posed by supervised machine learning techniques applied to problems that occur in the context of string theory. In particular, we have shown that the minimum volume V_{min} for Sasaki-Einstein 5-manifolds corresponding to toric Calabi-Yau 3-folds can be expressed by just 3 features of the associated toric diagrams Δ with an R^2-score ≥ 0.98. These three features are the area A of Δ, the number of vertices V in Δ, and the number of internal points in the factor $n = 3$ enlarged toric diagram Δ_3.

The simultaneous maximization of the R^2-score and the minimization of the number surviving parameters in the candidate function for $y = 1/V_{\text{min}}$ by varying the regularization strength given by the regularization parameter α, the proposed regularized regression models in this work give far more presentable, interpretable and explainable results than our previous work in [2]. Above all, as suggested in Fig. 9, the candidate formulas for the minimum volumes of toric Calabi-Yau 3-folds obtained in this study are concise enough to facilitate the examination of why some toric Calabi-Yau 3-folds are associated with minimum volumes that are more challenging to predict than those of certain other toric Calabi-Yau 3-folds. We plan to report on these investigations in the near future. We foresee that the application of regularization schemes to other supervised machine learning applications in string theory will open up equally promising research opportunities in the future.

ACKNOWLEDGMENTS

R. K.-S. would like to thank the Simons Center for Geometry and Physics at Stony Brook University, the City University of New York Graduate Center, the Institute for Basic Science Center for Geometry and Physics, as well as the Kavli Institute for the Physics and Mathematics of the Universe for hospitality during various stages of this work. R. K.-S. is supported by a Basic Research Grant of the National Research Foundation of Korea (No. NRF-2022R1F1A1073128). R. K.-S. is also supported by a Start-up Research Grant for new faculty at UNIST (No. 1.210139.01), UNIST AI Incubator Grant No. 1.230038.01 and UNIST UBSI Grants No. 1.230168.01 and No. 1.230078.01, as well as an Industry Research Project No. 2.220916.01 funded by Samsung SDS in Korea. He is also partly supported by the BK21 Program ("Next Generation Education Program for Mathematical Sciences" No. 4299990414089) funded by the Ministry of Education in Korea and the National Research Foundation of Korea (NRF).

[1] Y.-H. He, Deep-learning the landscape, arXiv:1706.02714.
[2] D. Krefl and R.-K. Seong, Machine learning of Calabi-Yau volumes, Phys. Rev. D **96**, 066014 (2017).
[3] F. Ruehle, Evolving neural networks with genetic algorithms to study the string landscape, J. High Energy Phys. 08 (2017) 038.
[4] J. Carifio, J. Halverson, D. Krioukov, and B. D. Nelson, Machine learning in the string landscape, J. High Energy Phys. 09 (2017) 157.
[5] A. Cole, A. Schachner, and G. Shiu, Searching the landscape of flux vacua with genetic algorithms, J. High Energy Phys. 11 (2019) 045.
[6] A. Cole, G. J. Loges, and G. Shiu, Interpretable phase detection and classification with persistent homology, in *34th Conference on Neural Information Processing Systems* (2020), 12, arXiv:2012.00783.
[7] J. Halverson, A. Maiti, and K. Stoner, Neural networks and quantum field theory, Mach. Learn. Sci. Tech. **2**, 035002 (2021).
[8] S. Gukov, J. Halverson, F. Ruehle, and P. Sułkowski, Learning to unknot, Mach. Learn. Sci. Tech. **2**, 025035 (2021).
[9] S. Abel, A. Constantin, T. R. Harvey, and A. Lukas, Evolving heterotic gauge backgrounds: Genetic algorithms versus reinforcement learning, Fortschr. Phys. **70**, 2200034 (2022).

[10] S. Krippendorf, R. Kroepsch, and M. Syvaeri, Revealing systematics in phenomenologically viable flux vacua with reinforcement learning, arXiv:2107.04039.
[11] A. Cole, S. Krippendorf, A. Schachner, and G. Shiu, Probing the structure of string theory vacua with genetic algorithms and reinforcement learning, in *35th Conference on Neural Information Processing Systems* (2021), 11, arXiv:2111.11466.
[12] P. Berglund, Y.-H. He, E. Heyes, E. Hirst, V. Jejjala, and A. Lukas, New Calabi-Yau manifolds from genetic algorithms, arXiv:2306.06159.
[13] M. Demirtas, J. Halverson, A. Maiti, M. D. Schwartz, and K. Stoner, Neural network field theories: Non-Gaussianity, actions, and locality, Mach. Learn. Sci. Tech. **5**, 015002 (2024).
[14] K. Bull, Y.-H. He, V. Jejjala, and C. Mishra, Machine learning CICY threefolds, Phys. Lett. B **785**, 65 (2018).
[15] V. Jejjala, A. Kar, and O. Parrikar, Deep learning the hyperbolic volume of a knot, Phys. Lett. B **799**, 135033 (2019).
[16] C. R. Brodie, A. Constantin, R. Deen, and A. Lukas, Machine learning line bundle cohomology, Fortschr. Phys. **68**, 1900087 (2020).
[17] Y.-H. He and A. Lukas, Machine learning Calabi-Yau four-folds, Phys. Lett. B **815**, 136139 (2021).
[18] H. Erbin and R. Finotello, Machine learning for complete intersection Calabi-Yau manifolds: A methodological study, Phys. Rev. D **103**, 126014 (2021).
[19] V. Anagiannis and M. C. N. Cheng, Entangled q-convolutional neural nets, Mach. Learn. Sci. Tech. **2**, 045026 (2021).
[20] M. Larfors, A. Lukas, F. Ruehle, and R. Schneider, Numerical metrics for complete intersection and Kreuzer–Skarke Calabi–Yau manifolds, Mach. Learn. Sci. Tech. **3**, 035014 (2022).
[21] S. Krippendorf and M. Syvaeri, Detecting symmetries with neural networks, arXiv:2003.13679.
[22] D. S. Berman, Y.-H. He, and E. Hirst, Machine learning Calabi-Yau hypersurfaces, Phys. Rev. D **105**, 066002 (2022).
[23] J. Bao, Y.-H. He, and E. Hirst, Neurons on amoebae, J. Symb. Comput. **116**, 1 (2022).
[24] R.-K. Seong, Unsupervised machine learning techniques for exploring tropical coamoeba, brane tilings and Seiberg duality, Phys. Rev. D **108**, 106009 (2023).
[25] D. Martelli, J. Sparks, and S.-T. Yau, Sasaki-Einstein manifolds and volume minimisation, Commun. Math. Phys. **280**, 611 (2008).
[26] D. Martelli, J. Sparks, and S.-T. Yau, The geometric dual of a-maximisation for toric Sasaki- Einstein manifolds, Commun. Math. Phys. **268**, 39 (2006).
[27] W. Fulton, *Introduction to Toric Varieties*, Annals of Mathematics Studies (Princeton University Press, Princeton, NJ, 1993).
[28] N. C. Leung and C. Vafa, Branes and toric geometry, Adv. Theor. Math. Phys. **2**, 91 (1998).
[29] B. R. Greene, String theory on Calabi-Yau manifolds, in *Theoretical Advanced Study Institute in Elementary Particle Physics (TASI 96): Fields, Strings, and Duality* (1996), 6, pp. 543–726, arXiv:hep-th/9702155.
[30] M. R. Douglas, B. R. Greene, and D. R. Morrison, Orbifold resolution by D-branes, Nucl. Phys. **B506**, 84 (1997).
[31] E. Witten, Anti-de Sitter space and holography, Adv. Theor. Math. Phys. **2**, 253 (1998).
[32] I. R. Klebanov and E. Witten, Superconformal field theory on three-branes at a Calabi-Yau singularity, Nucl. Phys. **B536**, 199 (1998).
[33] M. R. Douglas and G. W. Moore, D-branes, quivers, and ALE instantons, arXiv:hep-th/9603167.
[34] A. E. Lawrence, N. Nekrasov, and C. Vafa, On conformal field theories in four-dimensions, Nucl. Phys. **B533**, 199 (1998).
[35] B. Feng, A. Hanany, and Y.-H. He, D-brane gauge theories from toric singularities and toric duality, Nucl. Phys. **B595**, 165 (2001).
[36] B. Feng, A. Hanany, and Y.-H. He, Phase structure of D-brane gauge theories and toric duality, J. High Energy Phys. 08 (2001) 040.
[37] J. M. Maldacena, The large N limit of superconformal field theories and supergravity, Adv. Theor. Math. Phys. **2**, 231 (1998).
[38] D. R. Morrison and M. R. Plesser, Nonspherical horizons. 1., Adv. Theor. Math. Phys. **3**, 1 (1999).
[39] B. S. Acharya, J. M. Figueroa-O'Farrill, C. M. Hull, and B. J. Spence, Branes at conical singularities and holography, Adv. Theor. Math. Phys. **2**, 1249 (1999).
[40] K. A. Intriligator and B. Wecht, The exact superconformal R symmetry maximizes a, Nucl. Phys. **B667**, 183 (2003).
[41] A. Butti and A. Zaffaroni, R-charges from toric diagrams and the equivalence of a- maximization and Z-minimization, J. High Energy Phys. 11 (2005) 019.
[42] A. Butti and A. Zaffaroni, From toric geometry to quiver gauge theory: The equivalence of a-maximization and Z-minimization, Fortschr. Phys. **54**, 309 (2006).
[43] S. S. Gubser, Einstein manifolds and conformal field theories, Phys. Rev. D **59**, 025006 (1999).
[44] M. Henningson and K. Skenderis, The holographic Weyl anomaly, J. High Energy Phys. 07 (1998) 023.
[45] S. Benvenuti, B. Feng, A. Hanany, and Y.-H. He, Counting BPS operators in gauge theories: Quivers, syzygies and plethystics, J. High Energy Phys. 11 (2007) 050.
[46] B. Feng, A. Hanany, and Y.-H. He, Counting gauge invariants: The Plethystic program, J. High Energy Phys. 03 (2007) 090.
[47] Y.-H. He, R.-K. Seong, and S.-T. Yau, Calabi–Yau volumes and reflexive polytopes, Commun. Math. Phys. **361**, 155 (2018).
[48] C.-F. Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae* (Henricus Dieterich, Göttingen, 1823).
[49] R. A. Fisher, On the mathematical foundations of theoretical statistics, Phil. Trans. R. Soc. A **222**, 309 (1922).
[50] W. Mendenhall, T. Sincich, and N. S. Boudreau, *A Second Course in Statistics: Regression Analysis* (Prentice Hall Upper Saddle River, NJ, 2003), Vol. 6.
[51] D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, Cambridge, England, 2009).
[52] J. D. Jobson, *Applied Multivariate Data Analysis: Regression and Experimental Design* (Springer Science & Business Media, New York, 2012).

[53] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).
[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. **60**, 84 (2012).
[55] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).
[56] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. **61**, 85 (2015).
[57] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, Nature (London) **323**, 533 (1986).
[58] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2009), Vol. 2.
[59] A. Tikhonov, Regularization of incorrectly posed problems, Sov. Math. Dokl. **1624** (1963).
[60] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B **58**, 267 (1996).
[61] D. Martelli and J. Sparks, Toric geometry, Sasaki-Einstein manifolds and a new infinite class of AdS/CFT duals, Commun. Math. Phys. **262**, 51 (2006).
[62] S. Benvenuti, S. Franco, A. Hanany, D. Martelli, and J. Sparks, An infinite family of superconformal quiver gauge theories with Sasaki-Einstein duals, J. High Energy Phys. 06 (2005) 064.
[63] S. Benvenuti and M. Kruczenski, From Sasaki-Einstein spaces to quivers via BPS geodesics: L**p,q—r, J. High Energy Phys. 04 (2006) 033.
[64] A. Butti, D. Forcella, and A. Zaffaroni, The dual superconformal theory for L**pqr manifolds, J. High Energy Phys. 09 (2005) 018.
[65] S. Franco, A. Hanany, K. D. Kennaway, D. Vegh, and B. Wecht, Brane dimers and quiver gauge theories, J. High Energy Phys. 01 (2006) 096.
[66] A. Hanany and K. D. Kennaway, Dimer models and toric diagrams, arXiv:hep-th/0503149.
[67] S. Franco, A. Hanany, D. Martelli, J. Sparks, D. Vegh, and B. Wecht, Gauge theories from toric geometry and brane tilings, J. High Energy Phys. 01 (2006) 128.
[68] R. Kenyon, An introduction to the dimer model, arXiv:math/0310326.
[69] P. Kasteleyn, Graph theory and crystal physics, Graph theory and theoretical physics (1967), 43.
[70] F. Hirzebruch, *Singularities and Exotic Spheres* (Societe Mathematic de France, 1968).
[71] E. Brieskorn, Beispiele zur differentialtopologie von singularitäten, Inventiones Mathematicae **2**, 1 (1966).
[72] E. Witten, Phases of $N = 2$ theories in two dimensions, Nucl. Phys. **B403**, 159 (1993).
[73] A. Butti, D. Forcella, A. Hanany, D. Vegh, and A. Zaffaroni, Counting chiral operators in quiver gauge theories, J. High Energy Phys. 11 (2007) 092.
[74] A. Hanany and A. Zaffaroni, The master space of supersymmetric gauge theories, Adv. High Energy Phys. **2010**, 427891 (2010).
[75] D. Forcella, A. Hanany, Y.-H. He, and A. Zaffaroni, The master space of $N = 1$ gauge theories, J. High Energy Phys. 08 (2008) 012.
[76] D. Forcella, A. Hanany, Y.-H. He, and A. Zaffaroni, Mastering the master space, Lett. Math. Phys. **85**, 163 (2008).
[77] P. Pouliot, Molien function for duality, J. High Energy Phys. 01 (1999) 021.
[78] N. Seiberg, Electric—magnetic duality in supersymmetric nonAbelian gauge theories, Nucl. Phys. **B435**, 129 (1995).
[79] C. E. Beasley and M. Ronen Plesser, Toric duality is Seiberg duality, J. High Energy Phys. 12 (2001) 001.
[80] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[81] G. Pick, Geometrisches zur zahlenlehre, Sitzenber. Lotos (Prague) **19**, 311 (1899).
[82] P. Berglund, B. Campbell, and V. Jejjala, Machine learning Kreuzer-Skarke Calabi-Yau threefolds, arXiv:2112.09117.
[83] K. Hori and C. Vafa, Mirror symmetry, arXiv:hep-th/0002222.
[84] B. Feng, Y.-H. He, K. D. Kennaway, and C. Vafa, Dimer models from mirror symmetry and quivering amoebae, Adv. Theor. Math. Phys. **12**, 489 (2008).
[85] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis* (John Wiley & Sons, New York, 2021).
[86] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics **12**, 55 (1970).
[87] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B **67**, 301 (2005).