Letter

Machine learning Sasakian and G_2 topology on contact Calabi-Yau 7-manifolds

Daattavya Aggarwal [a], Yang-Hui He [b,c,d,e], Elli Heyes [b,c], Edward Hirst [f,*],

Henrique N. Sá Earp [g], Tomás S.R. Silva [g]

[a] Department of Computer Science and Technology, University of Cambridge, CB3 0FD, UK
[b] Department of Mathematics, City, University of London, EC1V 0HB, UK
[c] London Institute for Mathematical Sciences, Royal Institution, London, W1S 4BS, UK
[d] Merton College, University of Oxford, OX1 4JD, UK
[e] School of Physics, NanKai University, Tianjin, 300071, P.R. China
[f] Centre for Theoretical Physics, Queen Mary, University of London, E1 4NS, UK
[g] Institute of Mathematics, Statistics and Scientific Computing, University of Campinas (Unicamp), 13083-859, Brazil

A R T I C L E   I N F O

A B S T R A C T

We propose a machine learning approach to study topological quantities related to the Sasakian and G_2-geometries of contact Calabi-Yau 7-manifolds. Specifically, we compute datasets for certain Sasakian Hodge numbers and for the Crowley-Nördstrom invariant of the natural G_2-structure of the 7-dimensional link of a weighted projective Calabi-Yau 3-fold hypersurface singularity, for 7549 of the 7555 possible $\mathbb{P}^4(\mathbf{w})$ projective spaces. These topological quantities are then machine learnt with high performance scores, where learning the Sasakian Hodge numbers from the $\mathbb{P}^4(\mathbf{w})$ weights alone, using both neural networks and a symbolic regressor which achieve R^2 scores of 0.969 and 0.993 respectively. Additionally, properties of the respective Gröbner bases are well-learnt, leading to a vast improvement in computation speeds which may be of independent interest. The data generation and analysis further induced novel conjectures to be raised.

1. Introduction

Motivation Since its inception in 2017 [1–4], the study of Calabi-Yau (CY) manifolds in the context of string theory compactifications, using machine learning techniques has flourished, encompassing a wide array of investigations. Notably, these methods have been employed to predict Hodge numbers [5–9], learn Ricci-flat Calabi-Yau metrics [10–14], forecast line bundle cohomologies [15], generate new Calabi-Yau manifolds [16], and uncover volume bounds on Sasaki-Einstein manifolds [17]. Furthermore, machine learning techniques have found various other applications in geometry and physics [18–28]. For an extensive review, see [29,30]. As important as Calabi-Yau compactification is to string theory, 7-manifolds of holonomy G_2 are crucial to M-theory compactification [31,32]. Nonetheless, there have yet been no successful applications of machine learning in the context of G_2-geometry, let alone on compact manifolds with such special holonomy. This is no doubt due to the scarcity of dedicated databases for G_2-manifolds, which in turn

reflects the difficulty in describing (torsion-free) G_2-manifolds systematically in terms of algebraic discrete data.

While the holonomy reduction to G_2 amounts to a difficult non-linear PDE, pragmatically it may be relaxed in a number of ways, by considering the fundamental notion of a G_2-*structure*: a non-degenerate 3-form φ which induces a so-called G_2-*metric* g_φ; its failure to give rise to a metric with $\mathrm{Hol}(g_\varphi) \subset G_2$ is encoded by its *full torsion tensor* $T := \nabla^{g_\varphi}\varphi$. A G_2-structure is *closed* if $d\varphi = 0$ and *coclosed* if $d\psi = 0$, where $\psi := *_\varphi \varphi$, and the torsion-free condition $T = 0$ is equivalent to φ being both closed and coclosed. We propose therefore to work with coclosed G_2-structures on certain *contact Calabi-Yau (cCY)* 7-manifolds, which are closely related to the weighted projective Calabi-Yau 3-folds famously studied in [33].

Despite their unsuitability to M-theory, torsionful G_2 structures retain relevance in the context of (3+7)-dimensional heterotic supergravity with flux, as demonstrated by [34–36]. Indeed, as shown by [37], one can explicitly solve the corresponding Strominger system on cCY

* Corresponding author.
  E-mail address: e.hirst@qmul.ac.uk (E. Hirst).

7-manifolds, by way of coclosed G_2-structures together yielding nontrivial scalar and G_2-instanton gauge fields, with constant dilaton, as well as an H-flux with prescribed Chern-Simons defect, in accordance to the 'anomaly-free' condition referred to as the heterotic Bianchi identity.

Topological invariants of CY links   Contact Calabi-Yau manifolds were introduced by Tomassini and Vezzoni in [38], and they consist of Sasakian manifolds endowed with a closed basic complex volume form, which is 'transversally holomorphic' in the sense of foliations. It was shown in [39] that such a manifold carries naturally a coclosed G_2-structure.

A special class of such structures arises from Calabi-Yau links, which were first discussed from the perspective of G_2 topology in [40]. A 7-dimensional weighted link K_f is obtained as the intersection of a possibly small $S^9 \subset \mathbb{C}^5$ with a weighted homogeneous affine variety (defined by the zero locus of the polynomial f) having an isolated singularity at the origin. Milnor showed that such links are 2-connected compact smooth manifolds, indeed K_f is the total space of a Hopf S^1-bundle over a (weighted) projective 3-orbifold in $\mathbb{P}^4(\mathbf{w})$, for appropriate choices of polynomial degree and weighted $\mathbb{C}^\times$-action, see §2. Interestingly, the dataset of possible weights that admit these CY 3-folds consists of the 7555 cases classified in [33]. Therefore, we pursue the construction of a Calabi-Yau link for each of these weight systems, computing the following two types of topological invariants.

From the perspective of Sasakian topology, the (basic) Hodge numbers $h^{p,q}$ can be obtained as the dimensions of certain linear subspaces of the Milnor algebra $\mathbb{M}_f = \mathbb{C}[[z_1, \ldots, z_5]]/J_f$, defined by the corresponding Jacobian ideal of f [41]. We provide the first systematic computation of the Sasaki-Hodge numbers $\{h^{3,0}, h^{2,1}\}$ for this class of 7-dimensional CY links.

On the other hand, considering their G_2-topology, a CY link bounds an 8-dimensional Milnor fibre which smoothly extends the G_2-structure φ as a spinor field, hence it is possible to explicitly compute the Crowley-Nordström (CN) homotopy invariant $\nu(\varphi) \in \mathbb{Z}/48\mathbb{Z}$, introduced in [42]. Building upon the calculations first carried out in [40], we obtain an exhaustive dataset of ν-invariants for Calabi-Yau links.

Machine learning cCY topology   We analyse these two sets of topological data from a perspective similar to what has been done for Calabi-Yau manifolds [1]. In the standard Calabi-Yau case, the weights defining the ambient projective space are sufficient to uniquely determine the Calabi-Yau 3-fold's Hodge numbers, motivating ML of the known formulas from weights to Hodge numbers [43]. However, in the 7-dimensional Calabi-Yau link case, no such explicit formula is known, and one would initially expect the specific polynomial coefficients chosen to change the topology. Extending the ML techniques to these link invariants would establish existence of an approximate formula for Sasakian Hodge numbers from the weight information, from which ML interpretability techniques may be used to uncover its true form. This formula would provide new insights into Sasakian structures, as well as being dramatically quicker to compute, as well as open the door for their application on other related invariants, including for G_2-structures as motivated by their ML in this work.

We therefore extend previous work learning CY Hodge numbers from weights, to predicting Calabi-Yau link topological properties (namely Sasakian Hodge numbers and CN invariants). We find that, whilst the machine is able to learn the Sasakian Hodge number topology of these manifolds with high-performance measures, the same cannot be established for the CN invariant. The datasets of the weighted Calabi-Yau polynomials used in the link construction, with the computed Sasakian Hodge numbers and CN invariants, as well as the scripts used for analysis and machine learning, are available on `GitHub` [44].

This letter is organised as follows: in §2 we survey some background to contact Calabi-Yau manifolds, G_2-geometry, and machine learning;

in §3 we describe the methodology for the construction of the Calabi-Yau link data and perform relevant statistical analysis of the datasets of invariants; in §4 we present the results of the machine learning investigations; and we conclude in §5, discussing some future prospects.

2. Background

2.1. Calabi-Yau links

One may interpret structure group reductions on an odd-dimensional contact metric manifold (K^{2n+1}, η, ξ, g) as 'even-dimensional' structures 'transverse' with respect to a S^1-action along the fibres of a submersion $S^1 \to K \to V$. Here $\eta \in \Omega^1(K)$ denotes the contact form and $\xi \in \mathcal{X}(K)$ its (unit) dual Reeb field, such that $\eta(\xi) = 1$. Whenever clear from context, we will omit mention of the Riemannian metric g, for simplicity.

In particular, Sasakian geometry may be seen as transverse Kähler geometry, corresponding to the reduction of the transverse holonomy group to $U(n)$. These are equipped in addition with a transverse complex structure $J \in \text{End}(TK)$ such that $J \circ J = -I_{TK} + \eta \otimes \xi$, yielding a decomposition of forms by basic bi-degree, and a transverse symplectic form $\omega = d\eta \in \Omega^{1,1}(K)$, all of which satisfy suitable compatibility conditions; for more details see e.g. [45, §2] or the canonical reference [46]. Furthermore, Sasakian manifolds with special transverse holonomy $SU(n)$ are studied by Habib and Vezzoni [39, § 6.2.1]:

Definition 1. A Sasakian manifold $(K^{2n+1}, \eta, \xi, J, \Upsilon)$ is said to be a *contact Calabi–Yau manifold* (cCY) if Υ is a nowhere-vanishing transverse form of horizontal type $(n, 0)$, such that

$$\Upsilon \wedge \bar{\Upsilon} = (-1)^{\frac{n(n+2)}{2}} \omega^n \quad \text{and} \quad d\Upsilon = 0, \quad \text{with} \quad \omega = d\eta.$$

A polynomial $f \in \mathbb{C}[z_1, \ldots, z_{n+2}]$, for $n \geq 2$, is said to be *weighted homogeneous* of degree d with weight vector $\mathbf{w} = (w_1, \ldots, w_{n+2}) \in \mathbb{Z}_{>0}^{n+2}$, if it is homogeneous of order d with respect to the $\mathbb{C}^\times(\mathbf{w})$-action on $\mathbb{C}^{n+2}$

$$(t, z) \mapsto t \cdot z = (t^{w_1} z_1, \ldots, t^{w_{n+2}} z_{n+2}).$$

Such an f defines an affine variety

$$V_f = (f) = \{z \in \mathbb{C}^{n+2} \mid f(z) = 0\},$$

which, in general, admits a singularity at the origin.

Assuming that the origin is an isolated singularity, the intersection of V_f with a surrounding small hypersphere S_ε^{2n+3} is a compact smooth $(2n + 1)$-manifold $K_f = V_f \cap S_\varepsilon^{2n+3}$, the so-called weighted link of the singularity [47]. A weighted link K_f of degree d and weight w is a *Calabi-Yau link* if

$$d = \sum_{i=1}^{n+2} w_i, \tag{1}$$

which precisely guarantees the existence of a cCY structure on K_f. The dimension of the moduli space of these cCY structures is well-understood and discussed in §5.

2.2. Sasakian Hodge numbers of a CY link

The $\mathbb{C}^\times(\mathbf{w})$-action on $\mathbb{C}^{n+2}$ induces a contact-metric S^1-action on K_f. It admits finitely many distinct isotropy subgroups, contained in some finite subgroup $\Gamma \subset S^1$, so that K_f admits a double fibration over a projective n-orbifold $V \subset \mathbb{P}^{n+1}(\mathbf{w})$,

$$\pi : K_f \longrightarrow K_f/\Gamma \longrightarrow K_f/S^1 = (V_f \setminus \{0\})/\mathbb{C}^\times := V_f^*$$

The following key theorem allows us to compute certain mixed Hodge numbers $h^{p,q}(K_f)$ from the dimensions of the primitive cohomology groups $H_0^n(V_f^*)$, for $p + q = n$, which in turn can be obtained from the

Milnor algebra. A brief survey of Sasakian Hodge numbers can be found in Appendix A.

Theorem 2 *([41, Theorem 1.2], [48,49]). Let f be a $\mathbf{w}$-homogeneous polynomial on $\mathbb{C}^n$ of degree d. Given $p + q = n$, let $\ell = (p + 1)d - \sum_i w_i$, and denote by $(\mathbb{M}_f)_\ell$ the linear subspace of the Milnor algebra consisting of degree ℓ elements.*

$$h^{p,q}(K_f) = \dim_{\mathbb{C}}(\mathbb{M}_f)_\ell.$$

When (1) *is satisfied, i.e. K_f is a Calabi-Yau link, the condition reduces to $\ell = pd$.*

Finally, Moriyama expresses the dimension of the moduli space of cCY structures on a given 7-dimensional link K_f, in terms of the Sasakian Hodge numbers [50]:

$$\dim \mathfrak{M}_{CY}(K_f, \mathcal{F}_\xi) = b^3(K_f) + h^{1,1}(K_f) - 1. \tag{2}$$

In particular, the third Betti number b^3 is completely determined by $h^{2,1}$ and $h^{3,0}$:

$$(h^{3,0} + h^{2,1})(K_f) = \frac{1}{2}b^3(K_f), \tag{3}$$

which we have computed in this work. The remaining term is $h_S^{1,1}$, which is not calculable via Theorem 2, however may be accessible by other means. For instance, in the study of Calabi-Yau manifolds [33,51, 52], there is a well-established notion of homological mirror symmetry between Hodge numbers. We propose that if one could extend this to a notion of mirror symmetry among links, perhaps one could access $h^{1,1}$ for a link as being $h^{2,1}$ of the respective 'mirror'. For this dataset, we have enumerated the $h^{2,1}$'s *exhaustively*, so we could in principle know all the terms in (2) and subsequently the dimensions of the moduli space of cCY structures, at least for CY link mirror pairs.

2.3. The Crowley-Nordström invariant on cCY 7-manifolds

For an arbitrary closed 7-manifold with G_2-structure (Y^7, φ), Crowley and Nordström have defined a $\mathbb{Z}/48\mathbb{Z}$-valued homotopy invariant $\nu(\varphi)$, which is a combination of topological data from a compact coboundary 8-manifold with Spin(7)-structure (W^8, Ψ) extending (Y^7, φ), in the sense that $Y = \partial W$ and $\Psi|_Y = \varphi$:

$$\nu(\varphi) := \chi(W) - 3\sigma(W) \bmod 48, \tag{4}$$

where χ the real Euler characteristic and σ is the signature.

In particular $(n = 3)$, specialising §2.1 to real dimension 7, a contact Calabi–Yau structure naturally induces a coclosed G_2-structure (with symmetric torsion):

Proposition 3 *([39, Corollary 6.8]). Every cCY manifold $(K^7, \eta, \xi, J, \Upsilon)$ is an S^1-bundle $\pi : K \to V$ over a Calabi–Yau 3-orbifold (V, ω, Υ), with connection 1-form η and curvature*

$$d\eta = \omega, \tag{5}$$

and it carries a cocalibrated G_2-structure

$$\varphi := \eta \wedge \omega + \mathrm{Re}\,\Upsilon, \tag{6}$$

with torsion $d\varphi = \omega \wedge \omega$ and Hodge dual 4-form

$$\psi = * \varphi = \frac{1}{2}\omega \wedge \omega + \eta \wedge \mathrm{Im}\,\Upsilon.$$

It therefore makes sense, for further instance, to systematically study the invariants $\nu(\varphi)$ associated to $\mathbf{w}$-homogeneous polynomials f defining Calabi-Yau links K_f.

2.4. Weak R-equivalence of weighted polynomials

In the light of Theorem 2, we will observe in §3 that the Sasakian Hodge numbers and the natural CN invariant of a CY link depend only upon the Milnor algebra $\mathbb{M}_f$. The relation between the Milnor algebras of different weighted homogeneous polynomials was examined in [53], where the notion of R-equivalence is introduced:

Definition 4. Two weighted homogeneous polynomials f, g on $\mathbb{C}^n$ are R−*equivalent* if there exists a diffeomorphism $\psi : \mathbb{C}^n \mapsto \mathbb{C}^n$ such that $f \circ \psi = g$.

Theorem 2 in [53] gives a sufficient condition for R-equivalence between two such polynomials, as quoted below:

Theorem 5. *Let f, g be $\mathbf{w}$-homogeneous polynomials on $\mathbb{C}^n$ of degree d, such that $\mathcal{J}_f = \mathcal{J}_g$; then f is R-equivalent to g.*

However, our initial empirical observations (as detailed in §3.3) suggested that any homogeneous polynomial with the same weight vector (up to permutations), and no further singularities, has the same ℓ-degree subspaces of the Milnor algebra, up to linear isomorphism. This then implies that their respective Sasakian Hodge numbers and CN invariants will be the same. This motivates us to propose the following definition and conjecture.

Definition 6. Two weighted homogeneous polynomials f, g on $\mathbb{C}^n$ are said to be *weakly R-equivalent* if the respective ℓ-degree linear subspaces of their Milnor algebras are isomorphic, for each ℓ such that $p + q = n$, as in Theorem 2.

Conjecture 7. *Consider two weighted homogeneous polynomials f, g on $\mathbb{C}^n$ of same degree d; if their weight vectors $\mathbf{w}_f$ and $\mathbf{w}_g$ coincide (up to permutations), then f and g are weakly R-equivalent.*

The Conjecture is somewhat surprising, since it encompasses cases in which the Jacobian ideals $\mathcal{J}_f$ and $\mathcal{J}_g$ are non-isomorphic, and thus $\mathbb{M}_f$ and $\mathbb{M}_g$ are not equivalent. As we will see in §3, although certain steps in the algorithms to compute the Sasakian Hodge numbers and CN invariant involve the Gröbner basis, which is directly related to the Jacobian ideal, the results of these computations seem to depend only on the initial sets of weights of the $\mathbb{C}^\times$-action.

2.5. Machine learning

Aiming at an audience in the community of mathematics and theoretical physics, we provide a very brief introduction to neural networks, which is the architecture we use in our investigation [30,54,55]. We begin by introducing the *neuron*, the building block of any neural network. A neuron is a vertex in an oriented graph, which takes in a set of input data $\{x_i\}$ and produces a single numerical output y, by the following three steps:

1. First, each input x_i is multiplied by a weight W_i: $W_i x_i$.
2. Next, all the weighted inputs are summed and a bias b is added: $\sum_i W_i x_i + b$.
3. Finally, the sum is passed through a non-linear activation function which produces an output: $\hat{y} = \mathrm{act}(\sum_i W_i x_i + b)$.

ReLU is perhaps the most standard example of a non-linear activation function, it is defined

$$\mathrm{ReLU}(x) := \begin{cases} x & x > 0 \\ 0 & \textit{otherwise} \end{cases} \tag{7}$$

and is the activation function used in this work. A neural network is then simply a collection of neurons stratified in a series of layers, whereby the neurons in each layer are connected by edges to neurons in the previous and next layers.

The process of training a neural network starts with partitioning the dataset into *training data*, from which the network will learn, and *test data*, which is only used after training to evaluate the network's performance. The training process involves repeatedly calculating the 'error', which is some measure of the difference between the predicted model outputs and the true known outputs for the training data. During training, these weights and biases are stochastically updated in order to reduce this error measure. Computing the error requires a choice of loss function, for regression problems typically one uses either mean absolute error (MAE) or mean squared error (MSE)

$$
\begin{aligned}
\text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \\
\text{MSE} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,
\end{aligned}
\tag{8}
$$

where y_i and $\hat{y}_i$ are the true and predicted values, respectively, and N is the dataset size. The method by which we change the weights and biases to minimise the loss is called the *optimisation algorithm*, the simplest of which is stochastic gradient descent (SGD). There are more advanced optimisation methods that build on SGD, of which Adam [56] is a popular choice and is the particular optimisation we adopt.

For regression tasks, typical performance metrics include MAE (8) as well the R^2 score (9), which is defined as the proportion of the variance in the dependent variable that is predictable from the independent variable(s):

$$
R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad \in \quad (-\infty, 1],
\tag{9}
$$

where

$$
\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,
\tag{10}
$$

is the mean output. Therefore an R^2 score close to 1 means the regression model is a good fit, whereas a score close to 0 means the model is a poor fit. Additionally, despite this being a regression problem, we introduce a classification-inspired metric: Accuracy. We define this be the proportion of predictions within a fixed distance from the true value over the test data, where this fixed distance is defined in terms of a bound which is 0.05 times the range of the true values. This also evaluates in the range $[0,1]$, where a value of 1 indicates perfect learning.

Cross-validation is a method commonly used to get an unbiased evaluation of the learning, whereby the full dataset is shuffled and then split into k non-overlapping subsets. Each subset acts as the test dataset once, whilst the remaining $(k-1)$ subsets are combined to create the complement training dataset. k independent identical neural network models are then each trained on one of these training datasets, then evaluated on the complementary test dataset, with the evaluation scores recorded. The mean evaluation scores, with their standard errors, are then calculated and used to measure the model performance.

Continuous datasets may be quantitatively tested for correlations via the Product Moment Correlation Coefficient (PMCC), as linear method providing first order insight about potential dataset dependencies. This measure is defined

$$
\text{PMCC}(X, Y) := \frac{\mathbb{E}(X - \mu_X)\mathbb{E}(Y - \mu_Y)}{\sigma_x \sigma_Y} \in [-1, 1],
\tag{11}
$$

for random variables X, Y with respective means μ_X, μ_Y and standard deviations σ_x, σ_Y, and expectation values over the datasets taken with $\mathbb{E}(\cdot)$. It takes values in the range stated with perfect (anti-)correlation represented by PMCC of 1 (-1), and no correlation for PMCC of 0.

A more interpretable supervised learning method is symbolic regression [57]. In this method a basis of functions is assembled into an expression which is fit to training data. The basis used later in this work was $\{+, -, *, /\}$. These methods are genetic algorithms at their core, in that a population of expressions (represented as trees) are first initialised and then evaluated on the training data with respect to a standard loss (8) perturbed with a parsimony term which rewards simplicity in the expression (alike regularisation techniques in traditional ML preventing overfitting). The fittest individuals are then selected for cross-breeding between their expression trees, with subsequent mutation, to form the next generation of expressions. This process is then iterated to convergence, providing an array of candidate expressions which one can deduce information about the true functional form from.

3. Data generation & analysis

As previously stated, the Calabi-Yau 3-folds arising in the link construction are hypersurfaces in complex weighted projective space $\mathbb{P}^4(\mathbf{w})$. Such spaces are compact Fano manifolds (with positive curvature), constructed through identification of $\mathbb{C}^5$ with a weight vector of 5 entries. It was shown in [33], that the list of weight vector combinations which lead to unique weighted projective spaces whose anticanonical divisors are compact and Ricci-flat is finite, with $N = 7555$ cases.

For each $\mathbb{P}^4(\mathbf{w})$, any hypersurface in the anticanonical divisor class can be represented as a weighted homogeneous polynomial of degree $\sum_i w_i$. Throughout this class, there is freedom in the choice of complex coefficients for each of the monomial terms in the hypersurface's defining polynomial equation. Any choice of coefficients, such that the surface does not become more singular, defines a Calabi-Yau 3-fold. All of these will share the same Hodge numbers, but may otherwise be topologically distinct [33]. In addition, there is redundancy between choices of coefficient sets due to polynomial symmetries (such as coordinate transformations, coefficient normalisation, etc.), allowing multiple sets of coefficients to define the same 3-fold.

The dataset of Calabi-Yau links considered in this work was constructed using one Calabi-Yau from each of the respective 7555 $\mathbb{P}^4(\mathbf{w})$'s. In each case, the Calabi-Yau polynomial was first selected to have all monomial coefficients as 1. Physically, this may be interpreted as considering equivalent points on the Coulomb branches of the vacuum expectation value moduli spaces, when the Calabi-Yau manifolds are used for string compactification [58]. However 1484 out of the 7555 polynomial hypersurfaces intersected with singularities in the ambient space, leading to a higher-dimensional singularity structure on the links. To avoid this, for these 1484 cases other polynomials were sampled, with coefficients from $\{1, 2, 3, 4, 5\}$, until the singularity structure was exclusively the isolated singularity at the origin - as required for the link construction.[1]

To exemplify this process, consider the weight vector $\mathbf{w} = (22, 29, 49, 50, 75)$, whose degree $d = 225 \ (= \sum_i w_i)$ monomial basis has 7 terms:

$$
z_1^8 z_3, \ z_1^4 z_2^3 z_3, \ z_1 z_2^7, \ z_1 z_2 z_3 z_4 z_5, \ z_2 z_4^4, \ z_4^3 z_5, \ z_5^3.
$$

We thus initialise the Calabi-Yau polynomial equation to:

$$
0 = a_1 z_1^8 z_3 + a_2 z_1^4 z_2^3 z_3 + a_3 z_1 z_2^7 + a_4 z_1 z_2 z_3 z_4 z_5
$$

---

[1] Practically, the dimension of the singular locus of each Calabi-Yau polynomial was computed over a finite field of prime characteristic (101). Since this field reduction from the complex numbers cannot decrease the dimension of the singularity structure, where the dimension was 0 the polynomial was accepted. Where the observed singularity dimension was higher, a selection of other primes (251, 1993, 1997) were used to check for bad field reduction, where the dimension was 0 in any of those cases the polynomial was accepted as the increase in observed singularity dimension was due to this bad reduction. Where the singularity dimension did not decrease to 0, the polynomial was resampled until one with singularity dimension 0 was found (each time only 1 resample was required).
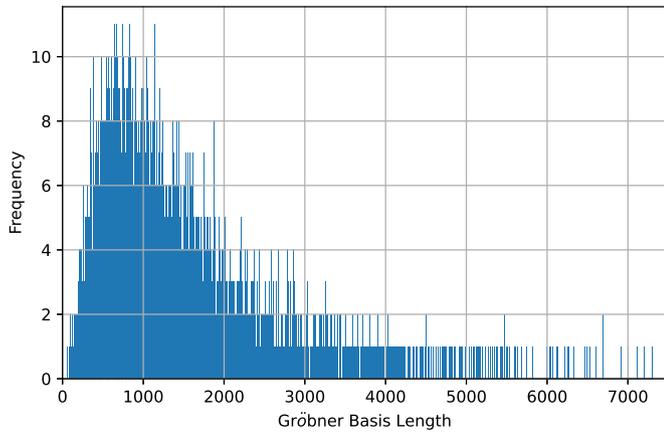
Fig. 1. Histogram of Gröbner basis lengths for the 7549 Calabi-Yau link constructions computed.

Fig. 2. Histogram of Sasakian $h^{2,1}$ values for the 7549 Calabi-Yau link constructions computed.

$$+ a_5 z_2 z_3^4 + a_6 z_4^3 z_5 + a_7 z_5^3,$$

for the complex coefficient vector $(a_1, \ldots, a_7)$. We set $a_1 = \cdots = a_7 = 1$, and check the singularity structure of the resulting hypersurface. In this case, the singular locus defined by this polynomial has dimension 0, which is the isolated singularity at the origin, there is hence no further singularity structure introduced. We therefore accept this Calabi-Yau 3-fold, adding it to our database for topological invariant computation (no further sampling of the a_i values is required).

For 7549 of these 7555 Calabi-Yau's selected in this way, the topological properties of the corresponding links were calculated. Namely, the Sasakian Hodge numbers $\{h^{3,0}, h^{2,1}\}$, from Theorem 2, and the CN invariant, from (4). It is worth emphasising that, since the list of weight vectors which lead to complex 3-dimensional Calabi-Yaus is finite, and since the topological invariants computed are conjectured to be identical for all Calabi-Yau polynomials with same weight vector (via Conjecture 7, and inspired by initial empirical observations exemplified in §3.3), the data generated for these 7555 manifolds would be *exhaustive* for this link construction.[2]

The polynomial generation and topological invariant computations were performed in `sagemath` [59], with the help of `macaulay2` [60] and `singular` [61]. Computation of each of the topological invariants required the respective Gröbner bases of the Calabi-Yau polynomials; these bases are notoriously expensive to compute with at worst doubly-exponential time complexity [62], and taking our High-Performance Computing cluster (HPC) ∼ 100,000 core hours. Hence, as a side product of these computational efforts, the Gröbner basis for a selection of the Calabi-Yau polynomials considered (one for each possible weight vector) is provided, along with the corresponding topological quantities, on this work's respective `GitHub`.

The distribution of the lengths of the Gröbner bases for 7549 out of 7555 Calabi-Yau polynomials is shown in Fig. 1. Due to the non-trivial connection between weights and basis length, and the importance of the basis length in determining whether invariant computation is even feasible, the prediction of Gröbner bases lengths was independently investigated, as detailed in §4.1.

---

[2] In this work, the topological data was computed for 7549 of the 7555, where the 6 remaining ((1, 1, 8, 19, 28), (1, 1, 9, 21, 32), (1, 1, 11, 26, 39), (1, 1, 12, 28, 42), (1, 6, 34, 81, 122), (1, 6, 40, 93, 140)) were the most computationally expensive and whose computation is left for future work. Since they represent such a small fraction of the dataset their omission should have negligible impact on learning performance.
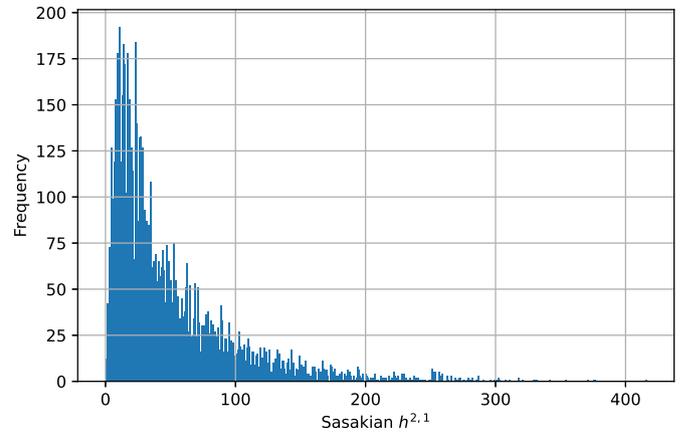
3.1. Sasakian Hodge numbers

As outlined in §2.2, the computation of the Hodge numbers $h^{3,0}$ and $h^{2,1}$ associated with each weighted homogeneous polynomial is done by the algorithmic implementation of the explicit formula in Theorem 2, from [41]:

Algorithm 1 Computation of Sasakian Hodge Numbers via Theorem 2.

Require: $f(z_1, \ldots, z_5)$, a homogeneous polynomial in $\mathbb{C}^5$.
Require: $\mathbf{w} = (w_1, \ldots, w_5)$, the weight vector associated with the polynomial f.
Ensure: $[h^{3,0}, h^{2,1}]$, the Sasakian Hodge Numbers associated to $(f, \mathbf{w})$

1: $A \leftarrow \mathbb{C}[z_1, \ldots, z_5]$
2: $d \leftarrow \deg(f)$ ($\triangleright$) w.r.t. $\mathbf{w}$
3: $J_f \leftarrow \left\langle \frac{\partial f}{\partial z_1}, \ldots, \frac{\partial f}{\partial z_5} \right\rangle$ ($\triangleright$) The Jacobian ideal of f
4: $K \leftarrow \text{GröbnerBasis}\left(\frac{A}{J_f}\right)$ ($\triangleright$) Basis of $\mathbb{M}_f$
5: $h^{3,0} = \#\{x \in K : \deg(x) = 4d - \sum_i w_i\}$ ($\triangleright$) $\dim(\mathbb{M}_f)_{4d - \sum w_i}$
6: $h^{2,1} = \#\{x \in K : \deg(x) = 3d - \sum_i w_i\}$ ($\triangleright$) $\dim(\mathbb{M}_f)_{3d - \sum w_i}$
7: **return** $[h^{3,0}, h^{2,1}]$

Step 4 of Algorithm 1 corresponds to a well-known super-exponential (hard) routine, the Gröbner basis generation [63,64]. In order to perform computations within a feasible time for all the 7555 polynomials,[3] we implemented a parallel version of Algorithm 1 using `sagemath` [59] and its built-in interface to `singular` [61], and executed the job on a HPC cluster.

The Sasakian $h^{3,0}$ values for the 7549 Calabi-Yau links computed all take value 1, matching the value known for all Calabi-Yau 3-folds, which corresponds to the unique holomorphic volume form. The Sasakian $h^{2,1}$ values range from 1 to 416, their frequency distribution is shown in Fig. 2. Despite the similar structure to the distribution of Gröbner basis lengths, we note that there is only a mild positive correlation, with PMCC ∼ 0.65.

Due to the aforementioned successes of ML in predicting the Hodge numbers of these Calabi-Yaus [1,8], it is natural to consider how this performance can extrapolate to Sasakian Hodge numbers, which will be addressed in §4.2. To compare directly the CY Hodge numbers with the Sasakian Hodge numbers of the links built from the same polynomials, a cross-plot of the respective $h^{2,1}$ values is given in Fig. 3. This plot shows that these topological invariants are strongly correlated (PMCC ∼ 0.99), and the Sasakian Hodge number is bounded above by the CY Hodge number – suggesting the following mathematical conjecture:

---

[3] Of which only 6 have been considered timed-out for this letter.
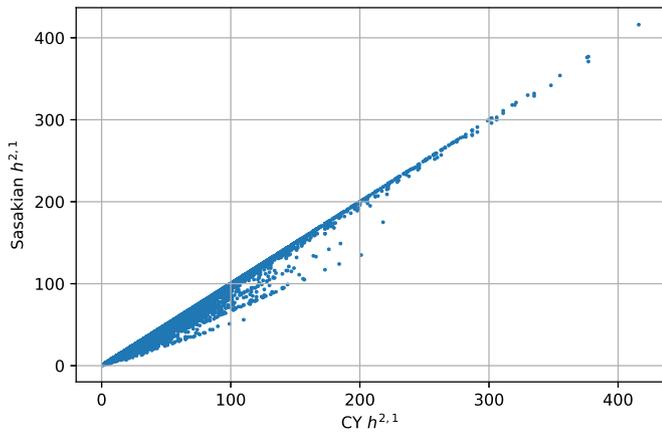
Fig. 3. Scatter graph of the Calabi-Yau complex threefold $h^{2,1}$ values against the Sasakian transverse $h^{2,1}$ values for the 7549 Calabi-Yau link constructions considered. For this data Sasakian $h^{2,1} \leq$ CY $h^{2,1}$, with 4198 cases satisfying the equality.

Conjecture 8. *The Sasakian Hodge number $h_S^{2,1}$ for a Calabi-Yau link is bounded above by the Hodge number $h_{CY}^{2,1}$ of the Calabi-Yau 3-fold built from the same $\mathbf{w}$-homogeneous polynomial:*

$$h_S^{2,1} \leq h_{CY}^{2,1} \, . \tag{12}$$

We note that the analogous bound also technically holds for $1 = h_S^{3,0} \leq h_{CY}^{3,0} = 1$; as it may well be the case for other yet uncomputed Sasakian Hodge numbers. Hence, from the successes in previous work on learning Calabi-Yau Hodge numbers [1,8], and this strong correlation with Sasakian Hodge numbers, the investigation into their ML prediction is well-motivated.

3.2. Crowley-Nördtrom invariant

To compute the CN invariant for polynomials in our dataset, we modify a procedure developed and described in [40] which utilises Steenbrink's Signature theorem. Let

$$\{z^\alpha : \alpha = (\alpha_1, ..., \alpha_{n+1}) \in I \subset \mathbb{Z}_{>0}^{n+1}\}$$

be a set of monomials in $\mathbb{C}[z_1, ..., z_{n+1}]$ representing a basis over $\mathbb{C}$ for its Milnor algebra $\mathbb{M}_f = \frac{\mathbb{C}[[z_1,...,z_{n+1}]]}{\partial f / \partial z_1, ..., \partial f / \partial z_{n+1}}$. For each $\alpha \in I$, define

$$l(\alpha) := \sum_{i=1}^{n+1} (\alpha_i + 1) \frac{w_i}{d}. \tag{13}$$

The CN invariant of a link was computed in [40] in terms of its degree and weights, along with the signature (μ_-, μ_0, μ_+) of the intersection form on $H^4(V_f, \mathbb{R})$:

$$\nu(\varphi) = \left(\frac{d}{w_1} - 1\right) ... \left(\frac{d}{w_5} - 1\right) - 3(\mu_+ - \mu_-) + 1 \tag{14}$$

Steenbrink [48] proved that the signature can be computed as follows:

$$\mu_+ = |\{\beta \in I : l(\beta) \notin \mathbb{Z}, \lfloor l(\beta) \rfloor \in 2\mathbb{Z}\}|$$

$$\mu_- = |\{\beta \in I : l(\beta) \notin \mathbb{Z}, \lfloor l(\beta) \rfloor \notin 2\mathbb{Z}\}|$$

$$\mu_0 = |\{\beta \in I : l(\beta) \in \mathbb{Z}\}|$$

In [40], this procedure was originally implemented as two separate scripts, one in `singular` and one in `MATHEMATICA` [65]. We improve upon this by combining those into a single `python` script. We are then able to take advantage of parallel processing, pooling, and the powerful computational resources of our HPC's to compute the CN invariant for 7549 out of the 7555 Calabi-Yau links.
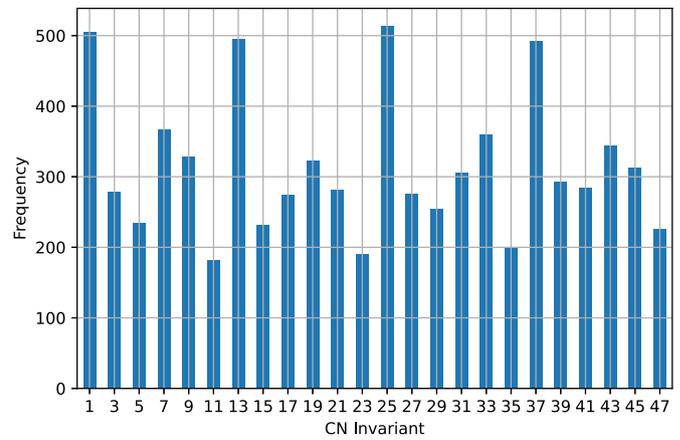
Fig. 4. Histogram of CN invariants for the 7549 Calabi-Yau link constructions computed.

The CN invariants computed fully span the range of possible values, which are odd integers from 1 to 47, cf. [40, Proposition 3.2]. Their frequency distribution is shown in Fig. 4, which exhibits an unexpected periodicity of 12 in the most populous invariant values (~ 500). In particular, we note the occurrence of CN invariants 27 and 35, where previous work had not identified examples in these topological classes [40]. Below we provide an explicit example of a Calabi-Yau polynomial that leads to a link in each of these classes (noting the repetition of the example considered earlier in §3):

CN : 27 ,

Weights : $(22, 29, 49, 50, 75)$, $\tag{15}$

Polynomial :

$$0 = z_1^8 z_3 + z_1^4 z_2^3 z_3 + z_1 z_2^7 + z_1 z_2 z_3 z_4 z_5 + z_2 z_3^4 + z_4^3 z_5 + z_5^3 ,$$

CN : 35 ,

Weights : $(31, 35, 36, 42, 108)$, $\tag{16}$

Polynomial :

$$0 = z_1^7 z_2 + z_1^2 z_2^2 z_3 z_4^2 + z_1 z_2 z_3^4 z_4 + z_1 z_2 z_3 z_4 z_5 + z_2^6 z_4 + z_3^7$$
$$+ z_3^4 z_5 + z_3 z_5^2 + z_4^6 .$$

3.3. Explicit weak R-equivalence

To corroborate the weak R-equivalence predicted in Conjecture 7, we show the observed behaviour for the previously selected example, as stated in (15).

In performing the checks of weak R-equivalence, we considered 10 permutations of the example weight system, and 50 polynomials per permutation (with general integer coefficients in the range (0,100), such that singular locus dimension is still 0; here, 100 was selected as it is the bound set by the prime 101 used for the coefficient ring characteristic). In each case, the computed CN invariant was $\nu = 27$, and $(h_S^{3,0}, h_S^{2,1}) = (1, 2)$. We observe that, while these values for the invariants were the same, the Gröbner basis lengths changed among different weight permutations (but were the same for different polynomials with the same weight system permutation). This behaviour was expected since the permutation of weights effectively amounts to a relabelling of the coordinates.

In addition to running these checks for the quoted example, the same procedure was repeated for 100 weight systems randomly selected from the database (selecting those with generally shorter polynomials, for computational efficiency), again considering 50 polynomials per weight system, and in all cases the weak R-equivalence was verified. Code to

run these checks in general scenarios is also made available in this article's `GitHub`

Below are two example weight permutations of (15), each with two respective Calabi-Yau polynomials. These were all included in the explicit checks above and hence lead to the same topological invariants.

$$\text{Weights} : [75, 22, 49, 29, 50] \,,$$

$$0 = z_1^3 + z_1 z_2 z_3 z_4 z_5 + z_1 z_5^3 + z_2^8 z_3 + z_2^4 z_4^3 z_5 + z_2 z_4^7 + z_3^4 z_4 \,,$$

$$0 = 48 z_1^3 + 49 z_1 z_2 z_3 z_4 z_5 + 6 z_1 z_5^3 + 71 z_2^8 z_3 + 35 z_2^4 z_4^3 z_5 + 29 z_2 z_4^7 + 25 z_3^4 z_4 \,.$$

$$\text{Weights} : [49, 22, 75, 50, 29] \,,$$

$$0 = 70 z_1^4 z_5 + 12 z_1 z_2^8 + 39 z_1 z_2 z_3 z_4 z_5 + 90 z_2^4 z_4 z_5^3 + 95 z_2 z_5^7 + 49 z_3^3 + 11 z_3 z_4^3 \,,$$

$$0 = 30 z_1^4 z_5 + 22 z_1 z_2^8 + 23 z_1 z_2 z_3 z_4 z_5 + 59 z_2^4 z_4 z_5^3 + 90 z_2 z_5^7 + 38 z_3^3 + 7 z_3 z_4^3 \,.$$

4. Machine learning

In order to investigate the efficacy of ML techniques to learn topological invariants of this dataset of Calabi-Yau links, NNs were chosen as the prototypical tool from supervised learning. Since the output invariants take a large range of values in each case, the NNs were set up for a regression-style problem. The NNs used had the same architecture in each case. They had neuron layer sizes of $(16, 32, 16)$, ReLU activation, and were trained on a MSE loss using an Adam optimiser. These layer sizes and the other hyperparameters were set after some heuristic tuning for the Gröbner basis ML, then used for the other investigations also for consistency. Each NN hence amounts to a map of the form:

$$\mathbb{R}^5 \xrightarrow{f_1} \mathbb{R}^{16} \xrightarrow{f_2} \mathbb{R}^{32} \xrightarrow{f_3} \mathbb{R}^{16} \xrightarrow{f_4} \mathbb{R}^1 \,, \tag{17}$$

such that each f_i acts via linear then non-linear action as $f(\mathbf{x}) = \text{ReLU}(\mathbf{W} \cdot \mathbf{x} + \mathbf{b})$.

In each case, the regression NNs were trained on 5 different partitions of the dataset into 80:20 train:test splits in accordance with cross-validation, to provide statistical error on the metrics used to assess learning performance. The NNs were implemented in `python` with the use of `scikit-learn` [66].

These NNs were trained to predict, from an input of the weight vectors, the respective Calabi-Yau link Gröbner basis length, Sasakian Hodge number $h^{2,1}$, and the CN invariant. The first two of these investigations are detailed in the subsequent sections, whilst the final investigation is briefly detailed here since this architecture could not learn to predict the CN invariant sufficiently well. Performance had R^2 value of ~ 0.004. Even reducing the problem to a binary classification between the two most populous classes ($\nu = 1$ and $\nu = 25$) did not lead to accuracies much above 0.5, indicating no significant learning and highlighting the highly non-trivial dependency of this invariant on the input polynomial and weight data, despite our computations showing the invariant was only dependent on the weight data in accordance with Conjecture (7).

4.1. Grobner basis length

By a significant margin, the computational bottleneck (in terms of both time and memory) of the invariant calculations was the generation of the Gröbner basis. Many initial runs failed from memory overload in this step for specific Calabi-Yau links. Through trial and error, we diverted our resources to allocate more computational power to the harder cases with larger Gröbner bases. However, it would have been substantially more efficient if we had approximately accurate predictions of which links would have required more computational power.

This problem again suits itself to ML, where a simple regression model can provide quick estimates of Gröbner basis length, and thus guide the subsequent allocation of computational resources. Previously, ML methods have been used to help optimise specific steps of the Gröbner basis algorithm [67,68], and decide when a Gröbner basis would be

useful [69]. Yet, only recently has ML been used to predict Gröbner basis properties directly, e.g. in [70] predictions of Gröbner basis length reached R^2 scores ~ 0.4 for binomial ideals of 2 terms.

For our specific construction the polynomial ideals have much more than 2 terms (with the longest polynomial having 680), and 5 terms per ideal for each partial derivative of the polynomial. Therefore it is prudent to investigate the performance of regression NNs in predicting Gröbner basis length for our Calabi-Yau links. These NNs used the same hyperparameters as previously specified in (17), with neuron layer sizes (16,32,16), ReLU activation, training on a MSE loss with an Adam optimiser on a 80:20 train:test data split. Performance measures for the learning were:

$$R^2 = 0.964 \pm 0.002 \,,$$

$$\text{MAE} = 122 \pm 2 \,, \tag{18}$$

$$\text{Accuracy} = 0.947 \pm 0.005 \,,$$

showing excellent performance, especially when comparing the MAE score to the Gröbner basis range of values, as shown in Fig. 1. We also note that whilst the number of monomial terms in a polynomial does correlate with Gröbner basis length, the correlation is not strong (PMCC ~ 0.66), and hence ML is much more useful in estimating the length of a Gröbner basis. These results corroborate the motivation to use ML in improving computational efficiency, particularly here in the application of calculating topological invariants; where perhaps the architectures are using that the data represents similar geometries to aid learning in this case. Inspired by this success, the authors hope to extend, in future work, the application of ML to learning more properties of Gröbner bases, as well as the basis elements directly.

4.2. Sasakian Hodge numbers

Extending work where NNs have showed surprising success in predicting Hodge numbers of Calabi-Yau manifolds, we now investigate their success in predicting the Sasakian Hodge number $h^{2,1}$, the computation of which was described in §3.1.

Using the same NN regressor architecture previously described in (17), with neuron layer sizes (16,32,16), ReLU activation, training on a MSE loss with an Adam optimiser on a 80:20 train:test data split. The $h^{2,1}$ values for the 7549 Calabi-Yau links computed were learnt with performance measures:

$$R^2 = 0.969 \pm 0.003 \,,$$

$$\text{MAE} = 5.53 \pm 0.22 \,, \tag{19}$$

$$\text{Accuracy} = 0.967 \pm 0.004 \,.$$

These results are equivalently strong and exemplify the efficacy of ML methods in predicting more subtle topological parameters.

Comparing the predicted $h^{2,1}$ outputs of a trained NN to the $h_{CY}^{2,1}$ values of the input weight system (instead of the intended $h_S^{2,1}$ values) produces lower performance scores across the dataset: $R^2 = 0.915$, MAE $= 9.88$, Accuracy $= 0.912$. This reassures that the NN's are learning the Sasakian topology instead of the correlated (and well learnt in previous work) Calabi-Yau base topology.

4.3. Sasakian Hodge symbolic regression

Motivated by the highly accurate regression results, one is led to expect these NN functions to be approximating a true relation between the Sasakian Hodge numbers and the weights used to define the Calabi-Yau links. In spirit, this could be a similar phenomenon to what was observed for the Poincaré polynomial, cf. [71].

To distil some mathematical insight about the function space of appropriate approximations for Sasakian Hodge numbers from weights, as equivalently and independently probed by the NN architectures, we

(a) Neural Network (b) Symbolic Regression
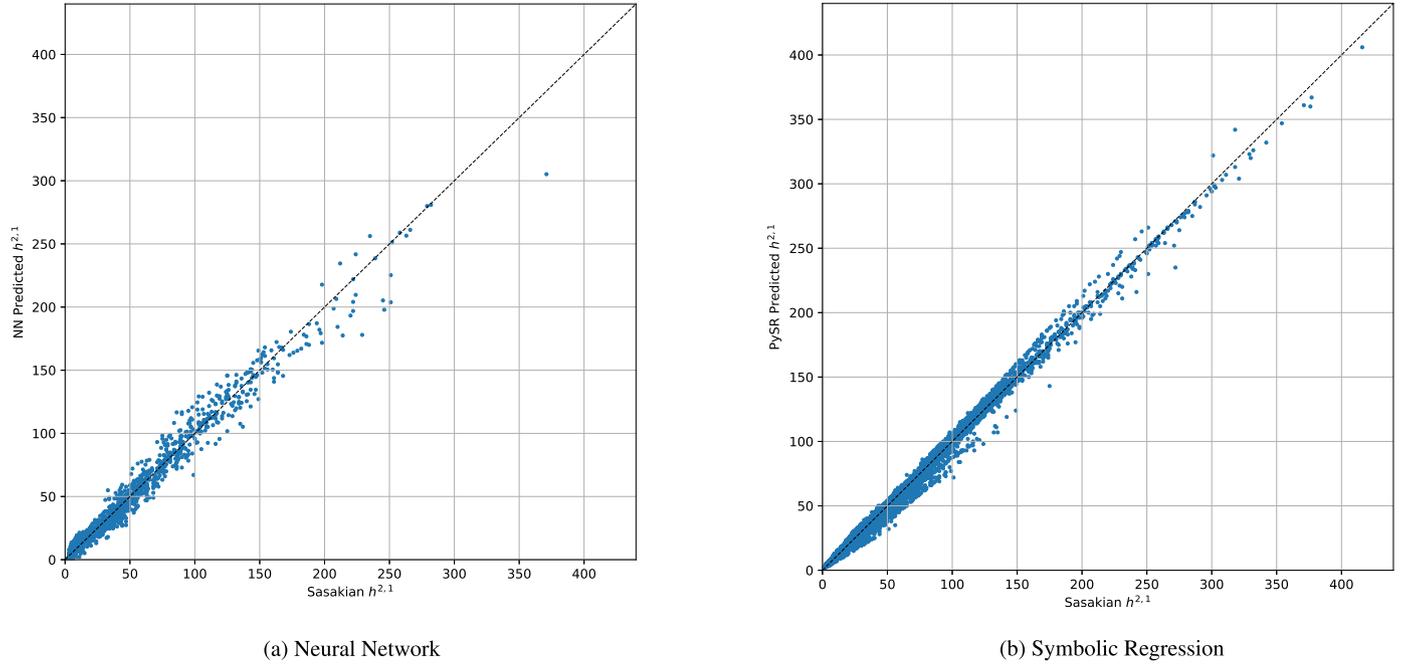
Fig. 5. ML architecture predictions of the $h_S^{2,1}$ values, against the true values, for the 7549 Calabi-Yau link constructions considered, from: (a) a trained NN; and (b) the symbolic regression best model of equation (20).

implement techniques of symbolic regression using `PySR` [72] to provide interpretable relations between inputs and outputs, which should help guide any investigation of this direct relation bypassing the Milnor algebra Gröbner basis computation. The scripts used for the symbolic regression analysis are also available at this work's `GitHub`. The (highest performing) equation on the independent test data (10% of the dataset) proposed by `PySR` to model Sasakian $h^{2,1}$ is presented in (20), achieving $R^2 \approx 0.99$ and MAE ≈ 2.6, exceeding the scores of the NN in (19).

$$h_{\mathrm{PySR}}^{2,1}(w_0, \ldots, w_4) = \frac{14.91 w_1 \left(w_0 w_4 + w_3 \left(w_0 + w_3 \right) \right)}{w_0 w_1 w_2 w_3} + \frac{10.02 w_2 w_3 \left(w_0 + w_4 + 0.77 \right)}{w_0 w_1 w_2 w_3} , \quad (20)$$

where this expression may shed some light on the structure of the true function.

The $h_S^{2,1}$ predicted value from (20) for each of the Calabi-Yau links considered is plotted against the true values in Fig. 5. Additionally the plot shows the equivalent predictions for a trained NN, where predictions are less accurate, particularly at higher invariant values.

4.4. Predicting the remaining invariants

As mentioned in §3, in this work the topological invariants have been computed for all but 6 of the 7555 weight systems, where HPC walltime has been successively exceeded due to the enormous computational cost. Practically the computations failed at walltimes of 1000 core hours using > 200 GB of RAM. However, as demonstrated in the previous sections, machine learning architectures have shown great success in modelling the prediction of these topological invariants. It therefore makes sense to apply instances of these very trained models to predict the invariant values for the remaining (currently computing) six weight systems. The prediction results are shown in Table 1.

Due to the excessive computation time for these remaining 6 weight systems, we expect their Gröbner basis lengths to make up the tail end of the respective distribution in Fig. 1. The extrapolation capabilities of the NN models were briefly tested by training the same NN architecture to predict Gröbner basis length from input weight system for 95% of the dataset pairs with shortest lengths, then testing on the remaining 5%

Table 1

ML architecture predictions of the well-learnt CY link properties for the remaining 6 weight systems. The Gröbner basis length (GBL) and the Sasakian $h_S^{2,1}$ values are predicted by the trained NN architectures or the symbolic regression (SR) equation.

Invariant	GBL	$h_S^{2,1}$	
Architecture	NN	NN	SR
(1, 1, 8, 19, 28)	1531	257	338
(1, 1, 9, 21, 32)	1623	272	377
(1, 1, 11, 26, 39)	1807	293	447
(1, 1, 12, 28, 42)	1879	299	476
(1, 6, 34, 81, 122)	3488	243	243
(1, 6, 40, 93, 140)	3770	265	272

(378 bases). The NN had lower performance scores of: $R^2 = 0.594$, MAE = 452, Accuracy = 0.484; highlighting the importance of interpretation of statistical methods for out-of-distribution predictions. However the predictions do well identify these test weight systems as having long bases, with the minimum predicted basis length on the test data being 2391, far exceeding the mean of the full dataset (1290). Whilst the NN models do predict large lengths for the remaining weight systems' Gröbner bases in Table 1, with values significantly above the average, they do not predict values exceeding the current highest value of 7299. These predictions are high, as is useful for our implementation but less likely to be exact.

Conversely, the Sasakian Hodge number values have been shown to only loosely correlate with the Gröbner basis length, hence there is only loose intuition that these Hodge numbers should be high. Our Conjecture 8 produces some bound on the Sasakian $h^{2,1}$ values, set by the respective Calabi-Yau $h^{2,1}$ values: 348, 387, 462, 491, 246, 275. Interestingly, all predictions from both the NN and the symbolic regression satisfy this bound as well. Furthermore, the better-performing symbolic regression model of (20) predicts Hodge numbers very close to this bound, which is sensible behaviour, as demonstrated by Fig. 3.

Despite the predictions between the NN and symbolic regressor in some cases being quite different, pragmatically mathematicians and physicists are often most interested in discerning when Hodge numbers are particularly low, such as 0 or 1, as indications of phenomena such as exactness, rigidity or unobstruction. Hence, these learnt mod-

els can very quickly provide confidence in whether this is the case, by making predictions which can test conjectures or corroborate theoretical expectations.

Our own experienced difficulty in computing the topological invariants for these final 6 remaining cases illustrates the potential of these machine learning models. Where direct computation is not feasible, ML methods can provide predictions for quantities of interest (such as our CY link topological invariants) with statistical confidence, providing invaluable insight to guide refinement of the computation, and further the progress of academic research.

5. Conclusion

In this work, real 7-dimensional Calabi-Yau links were constructed for a complex 3-dimensional Calabi-Yau manifold coming from 7549 of the 7555 complex 4-dimensional weighted projective spaces that admit them. It was observed, and conjectured, that any Calabi-Yau hypersurface with the correct singularity properties leads to the same Sasakian Hodge number $h^{2,1}$ and CN invariant; which (when computed for the final 6 evasive cases) will produce an exhaustive list of these invariants from the Calabi-Yau link construction.

The datasets of these invariants were statistically analysed, and NN regressors were used to successfully predict the respective Sasakian $h^{2,1}$ values from the ambient $\mathbb{P}^4(\mathbf{w})$ weights alone. The same architectures were not successful in predicting the CN invariant, but did show surprising success in predicting the length of the Gröbner basis from the weights. These regressors can hence be used to streamline future computation of Gröbner basis by informing on efficient computational resource allocation.

The exhaustive list of Calabi-Yau link data, as well as the python scripts used for their analysis and ML are made available at this work's corresponding GitHub.

Avenues for future work include studying whether we can obtain Sasakian structure for more general links made through general toric varieties rather than weighted projective spaces. As well as application of ML techniques to study other Gröbner basis properties to further streamline their computation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Code and Data available at: https://github.com/TomasSilva/MLcCY7 (also quoted in paper).

Appendix A. Sasakian Hodge decomposition

We establish some notation and elementary facts about the complexified tangent bundle, which are largely adapted from [73]. The

contact structure splits the tangent bundle as $TM = B \oplus N_\xi$, where $B = \ker(\eta)$ and N_ξ is the real line bundle spanned by the Reeb field ξ. The transverse complex structure Φ satisfies $(\Phi|_B)^2 = -1$, so the eigenvalues of the complexified operator $\Phi^{\mathbb{C}}$ are $\pm\sqrt{-1}$. The complexification $B_{\mathbb{C}} := B \otimes_{\mathbb{R}} \mathbb{C}$ splits as $B_{\mathbb{C}} = B^{1,0} \oplus B^{0,1}$, so we obtain a decomposition of direct sum of vector bundles

$$\Lambda^d(B_{\mathbb{C}})^* = \bigoplus_{i=0}^{d}(B^{i,d-i})^*.$$

This induces the decomposition of vector bundles

$$\Omega^d(M) = \left(\bigoplus_{i=0}^{d}\Omega_B^{i,d-i}(M)\right) \oplus \left(\eta \wedge \left(\bigoplus_{j=0}^{d-1}\Omega_B^{j,d-j-1}(M)\right)\right),$$

where $\Omega_B^{p,q}(M) := \Gamma(M, \Lambda^p(B_{\mathbb{C}})^* \otimes \Lambda^q(B_{\mathbb{C}})^*)$.

Let us briefly describe the transverse complex geometry on a Sasakian manifold (M, η, ξ, g, Φ), following [45] and references therein. Relatively to the Reeb foliation, the usual Hodge star induces a *transverse Hodge star* operator $*_T : \Omega_B^k(M) \to \Omega_B^{m-(k+1)}(M)$ by the formula

$$*_T(\beta) = (-1)^{m-1-k} * (\beta \wedge \eta). \tag{A.1}$$

Both operators are compatible, in the sense that

$$* \alpha = *_T \alpha \wedge \eta, \quad \forall \alpha \in \Omega_B^\bullet(M). \tag{A.2}$$

Now, acting on (p,q)-forms, we have well-defined operators

$$\partial_B : \Omega_B^{p,q} \to \Omega_B^{p+1,q} \text{ and } \bar{\partial}_B : \Omega_B^{p,q} \to \Omega_B^{p,q+1}. \tag{A.3}$$

In terms of the transverse Hodge star (A.1), the operators (A.3) have adjoints

$$\begin{aligned} \partial_B^* : \quad &\Omega_B^{p,q} \to \Omega_B^{p-1,q} & &\text{and} & \bar{\partial}_B^* : \quad &\Omega_B^{p,q} \to \Omega_B^{p,q-1} \\ &\partial_B^* := -*_T \partial_B *_T & & & &\bar{\partial}_B^* := -*_T \bar{\partial}_B *_T. \end{aligned} \tag{A.4}$$

An inner product on $\Omega^\bullet(M)$ is defined by:

$$(\alpha, \beta)_M = \int_M \alpha \wedge * \beta = \int_M \alpha \wedge *_T \beta \wedge \eta. \tag{A.5}$$

Defining $\Delta_{\bar{\partial}_B} := \bar{\partial}_B^* \bar{\partial}_B + \bar{\partial}_B \bar{\partial}_B^*$, $H^{p,q} := \ker \Delta_{\bar{\partial}_B} \subset \Omega_B^{p,q}$, the space H^k of harmonic k-forms on M has the following decomposition [41,74]:

$$H^k = \bigoplus_{p+q=k} H^{p,q}, \quad \text{with} \quad \bar{H}^{p,q} = H^{q,p}, \quad k \leq n. \tag{A.6}$$

References

[1] Y.-H. He, Deep-learning the landscape, arXiv:1706.02714, 6.2017.

[2] F. Ruehle, Evolving neural networks with genetic algorithms to study the string landscape, J. High Energy Phys. 2017 (8) (Aug. 2017), https://doi.org/10.1007/jhep08(2017)038.

[3] J. Carifio, J. Halverson, D. Krioukov, B.D. Nelson, Machine learning in the string landscape, J. High Energy Phys. 2017 (9) (Sep. 2017), https://doi.org/10.1007/jhep09(2017)157.

[4] D. Krefl, R.-K. Seong, Machine learning of Calabi-Yau volumes, Phys. Rev. D 96 (6) (Sep. 2017), https://doi.org/10.1103/physrevd.96.066014.

[5] K. Bull, Y.-H. He, V. Jejjala, C. Mishra, Getting CICY High, Phys. Lett. B 795 (2019) 700–706, https://doi.org/10.1016/j.physletb.2019.06.067, arXiv:1903.03113.

[6] Y.-H. He, A. Lukas, Machine learning Calabi-Yau four-folds, Phys. Lett. B 815 (2021) 136139, https://doi.org/10.1016/j.physletb.2021.136139, arXiv:2009.02544.

[7] H. Erbin, R. Finotello, Machine learning for complete intersection Calabi-Yau manifolds: a methodological study, Phys. Rev. D 103 (12) (Jun. 2021), https://doi.org/10.1103/physrevd.103.126014.

[8] D.S. Berman, Y.-H. He, E. Hirst, Machine learning Calabi-Yau hypersurfaces, Phys. Rev. D 105 (6) (2022) 066002, https://doi.org/10.1103/PhysRevD.105.066002, arXiv:2112.06350.

[9] B. Aslan, D. Platt, D. Sheard, Group invariant machine learning by fundamental domain projections, in: S. Sanborn, C. Shewmake, S. Azeglio, A. Di Bernardo, N. Miolane (Eds.), Proceedings of the 1st NeurIPS Workshop on Symmetry and Geometry in Neural Representations, in: Proceedings of Machine Learning Research, PMLR, vol. 197, 2023, pp. 181–218, https://proceedings.mlr.press/v197/aslan23a.html.

[10] A. Ashmore, Y.-H. He, B.A. Ovrut, Machine learning Calab-Yau metrics, Fortschr. Phys. 68 (9) (2020) 2000068, https://doi.org/10.1002/prop.202000068.

[11] L.B. Anderson, M. Gerdes, J. Gray, S. Krippendorf, N. Raghuram, F. Ruehle, Moduli-dependent Calabi-Yau and SU(3)-structure metrics from machine learning, J. High Energy Phys. 05 (2021) 013, https://doi.org/10.1007/JHEP05(2021)013, arXiv:2012.04656.

[12] V. Jejjala, D.K. Mayorga Pena, C. Mishra, Neural network approximations for Calabi-Yau metrics, J. High Energy Phys. 08 (2022) 105, https://doi.org/10.1007/JHEP08(2022)105, arXiv:2012.15821.

[13] M.R. Douglas, S. Lakshminarasimhan, Y. Qi, Numerical Calabi-Yau metrics from holomorphic networks, arXiv:2012.04797, 12.2020.

[14] M. Larfors, A. Lukas, F. Ruehle, R. Schneider, Learning size and shape of Calabi-Yau spaces, arXiv:2111.01436, 11.2021.

[15] D. Klaewer, L. Schlechter, Machine learning line bundle cohomologies of hypersurfaces in toric varieties, Phys. Lett. B 789 (2019) 438–443, https://doi.org/10.1016/j.physletb.2019.01.002.

[16] P. Berglund, Y.-H. He, E. Heyes, E. Hirst, V. Jejjala, A. Lukas, New Calabi-Yau manifolds from genetic algorithms, arXiv:2306.06159, 6.2023.

[17] M. Manko, An upper bound on the critical volume in a class of toric Sasaki-Einstein manifolds, arXiv:2209.14029, 9.2022.

[18] Y.-H. He, E. Hirst, T. Peterken, Machine-learning dessins d'enfants: explorations via modular and Seiberg-Witten curves, J. Phys. A 54 (7) (2021) 075401, https://doi.org/10.1088/1751-8121/abbc4f, arXiv:2004.05218.

[19] J. Bao, S. Franco, Y.-H. He, E. Hirst, G. Musiker, Y. Xiao, Quiver mutations, Seiberg duality and machine learning, Phys. Rev. D 102 (8) (2020) 086013, https://doi.org/10.1103/PhysRevD.102.086013, arXiv:2006.10783.

[20] J. Bao, O. Foda, Y.-H. He, E. Hirst, J. Read, Y. Xiao, F. Yagi, Dessins d'enfants, Seiberg-Witten curves and conformal blocks, J. High Energy Phys. 05 (2021) 065, https://doi.org/10.1007/JHEP05(2021)065, arXiv:2101.08843.

[21] J. Bao, Y.-H. He, E. Hirst, J. Hofscheier, A. Kasprzyk, S. Majumder, Hilbert series, machine learning, and applications to physics, Phys. Lett. B 827 (2022) 136966, https://doi.org/10.1016/j.physletb.2022.136966, arXiv:2103.13436.

[22] J. Bao, Y.-H. He, E. Hirst, Neurons on amoebae, J. Symb. Comput. 116 (2022) 1–38, https://doi.org/10.1016/j.jsc.2022.08.021, arXiv:2106.03695.

[23] J. Bao, Y.-H. He, E. Hirst, J. Hofscheier, A. Kasprzyk, S. Majumder, Polytopes and machine learning, arXiv:2109.09602, 9.2021.

[24] G. Arias-Tamargo, Y.-H. He, E. Heyes, E. Hirst, D. Rodriguez-Gomez, Brain webs for brane webs, Phys. Lett. B 833 (2022) 137376, https://doi.org/10.1016/j.physletb.2022.137376, arXiv:2202.05845.

[25] P.-P. Dechant, Y.-H. He, E. Heyes, E. Hirst, Cluster algebras: network science and machine learning, arXiv:2203.13847, 3.2022.

[26] S. Chen, Y.-H. He, E. Hirst, A. Nestor, A. Zahabi, Mahler measuring the genetic code amoebae, arXiv:2212.06553, 12.2022.

[27] M.-W. Cheung, P.-P. Dechant, Y.-H. He, E. Heyes, E. Hirst, J.-R. Li, Clustering cluster algebras with clusters, arXiv:2212.09771, 12.2022.

[28] A. Ashmore, Y.-H. He, E. Heyes, B.A. Ovrut, Numerical spectra of the Laplacian for line bundles on Calabi-Yau hypersurfaces, arXiv:2305.08901, 5.2023.

[29] J. Bao, Y.-H. He, E. Heyes, E. Hirst, Machine learning algebraic geometry for physics, arXiv:2204.10334, 4.2022.

[30] Y.-H. He, E. Heyes, E. Hirst, Machine learning in physics and geometry, arXiv:2303.12626, 3.2023.

[31] B. Acharya, E. Witten, Chiral fermions from manifolds of G_2 holonomy, arXiv:hep-th/0109152v1, 2001.

[32] B. Acharya, S. Gukov, M theory and singularities of exceptional holonomy manifolds, Phys. Rep. 392 (3) (2004) 121–189.

[33] P. Candelas, M. Lynker, R. Schimmrigk, Calabi-Yau manifolds in weighted P4, Nucl. Phys. B 341 (2) (1990) 383–402, https://doi.org/10.1016/0550-3213(90)90185-G, https://www.sciencedirect.com/science/article/pii/055032139090185G.

[34] X. de la Ossa, M. Larfors, E.E. Svanes, Exploring $SU(3)$ structure moduli spaces with integrable G_2 structures, Adv. Theor. Math. Phys. 19 (2015) 837–903, https://doi.org/10.4310/ATMP.2015.v19.n4.a5, arXiv:1409.7539.

[35] X. de la Ossa, M. Larfors, E.E. Svanes, The infinitesimal moduli space of heterotic G_2 systems, Commun. Math. Phys. 360 (2) (2018) 727–775, https://doi.org/10.1007/s00220-017-3013-8, arXiv:1704.08717.

[36] X. de la Ossa, M. Larfors, E.E. Svanes, Restrictions of heterotic G_2 structures and instanton connections, in: Nigel Hitchin's 70th Birthday Conference, 2017, arXiv:1709.06974.

[37] J. Lotay, H.S. Earp, The heterotic G_2 system on contact Calabi–Yau 7-manifolds, Trans. Amer. Math. Soc. Ser. B 10 (26) (2023) 907–943, https://doi.org/10.1090/btran/129.

[38] A. Tomassini, L. Vezzoni, Contact Calabi-Yau manifolds and special Legendrian submanifolds, arXiv:math/0612232, 2007.

[39] G. Habib, L. Vezzoni, Some remarks on Calabi-Yau and hyper-Kähler foliations, Differ. Geom. Appl. 41 (2015) 12–32, https://doi.org/10.1016/j.difgeo.2015.03.006.

[40] O. Calvo-Andrade, L. Rodríguez, H.N. Sá Earp, Gauge theory and G_2-geometry on Calabi-Yau links, Rev. Mat. Iberoam. 36 (6) (2020) 1753–1778, arXiv:1606.09271 [math.DG], https://www.ems-ph.org/journals/show_pdf.php?issn=0213-2230&vol=36&iss=6&rank=6.

[41] M. Itoh, Sasakian manifolds, Hodge decomposition and Milnor algebras, Kyushu J. Math. 58 (2004) 121–140, https://www.jstage.jst.go.jp/article/kyushujm/58/1/58_1_121/_article/-char/ja/.

[42] D. Crowley, J. Nordström, New invariants of G_2-structures, Geom. Topol. 19 (2015) 2949–2992.

[43] C. Vafa, String vacua and orbifolded L-G models, Mod. Phys. Lett. A 4 (1989) 1169, https://doi.org/10.1142/S0217732389001350.

[44] D. Aggarwal, Y.-H. He, E. Heyes, E. Hirst, H.N.S. Earp, T.S.R. Silva, https://github.com/TomasSilva/MLcCY7, 2023.

[45] L.E. Portilla, H.N. Sá Earp, Instantons on Sasakian 7-manifolds, Q. J. Math. (2023), https://academic.oup.com/qjmath/advance-article-pdf/doi/10.1093/qmath/haad011/49696708/haad011.pdf, https://doi.org/10.1093/qmath/haad011.

[46] C.P. Boyer, K. Galicki, Sasakian Geometry, Oxford Mathematical Monographs, Oxford University Press, Oxford, 2008.

[47] J. Milnor, Singular points of complex hypersurfaces, 1969.

[48] J. Steenbrink, Intersection form for quasi-homogeneous singularities, Compos. Math. 34 (2) (1977) 211–223.

[49] J.H.M. Steenbrink, Mixed Hodge structures associated with isolated singularities, https://doi.org/10.1090/PSPUM/040.2/713277, 1983.

[50] T. Moriyama, The moduli space of transverse Calabi-Yau structures on foliated manifolds, https://doi.org/10.18910/12555, 2009.

[51] L.J. Dixon, Some world sheet properties of superstring compactifications, on orbifolds and otherwise, in: Summer Workshop in High-Energy Physics and Cosmology, 1987.

[52] W. Lerche, C. Vafa, N.P. Warner, Chiral rings in N = 2 superconformal theories, Nucl. Phys. B 324 (1989) 427–474, https://doi.org/10.1016/0550-3213(89)90474-4.

[53] I. Ahmed, Weighted homogeneous polynomials with isomorphic Milnor algebras, J. Prime Res. Math. (2012).

[54] J.A. Anderson, An Introduction to Neural Networks, MIT Press, 1995.

[55] F. Ruehle, Data science applications to string theory, Phys. Rep. 839 (2020) 1–117, https://doi.org/10.1016/j.physrep.2019.09.005.

[56] D.P. Kingma, J. Ba Adam, A method for stochastic optimization, arXiv:1412.6980, 2017.

[57] J.R. Koza, Genetic programming as a means for programming computers by natural selection, Stat. Comput. 4 (2) (1994) 87–112, https://doi.org/10.1007/BF00175355.

[58] B. Feng, A. Hanany, Y.-H. He, D-brane gauge theories from toric singularities and toric duality, Nucl. Phys. B 595 (2001) 165–200, https://doi.org/10.1016/S0550-3213(00)00699-4, arXiv:hep-th/0003085.

[59] T.S. Developers, W. Stein, D. Joyner, D. Kohel, J. Cremona, B. Eröcal, SageMath, version 9.0, http://www.sagemath.org, 2020.

[60] D.R. Grayson, M.E. Stillman, Macaulay2, a software system for research in algebraic geometry, Available at http://www.math.uiuc.edu/Macaulay2/, 1992.

[61] W. Decker, G.-M. Greuel, G. Pfister, H. Schönemann, Singular 4-3-0 — a computer algebra system for polynomial computations, http://www.singular.uni-kl.de, 2022.

[62] T.W. Dubé, The structure of polynomial ideals and Gröbner bases, SIAM J. Comput. 19 (4) (1990) 750–773, https://doi.org/10.1137/0219053.

[63] M. Bardet, J.-C. Faugère, B. Salvy, On the complexity of the F5 Gröbner basis algorithm, J. Symb. Comput. 70 (2015) 49–70, https://doi.org/10.1016/j.jsc.2014.09.025, https://www.sciencedirect.com/science/article/pii/S0747717114000935.

[64] A. William, P. Loustaunau, An Introduction to Grobner Bases, American Mathematical, 1994.

[65] W. R. Inc., Mathematica, Version 13.1, champaign, IL, https://www.wolfram.com/mathematica, 2022.

[66] 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, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[67] D. Peifer, M. Stillman, D. Halpern-Leistner, Learning selection strategies in Buchberger's algorithm, arXiv:2005.01917, 2020.

[68] J. Mojsilović, D. Peifer, S. Petrović, Learning a performance metric of Buchberger's algorithm, Involve 16 (2) (2023) 227–248, https://doi.org/10.2140/involve.2023.16.227.

[69] Z. Huang, M. England, J.H. Davenport, L.C. Paulson, Using machine learning to decide when to precondition cylindrical algebraic decomposition with Groebner bases, https://doi.org/10.1109/synasc.2016.020, Sep. 2016.

[70] S. Jamshidi, E. Kang, S. Petrović, Predicting the cardinality of a reduced Gröbner basis, arXiv:2302.05364, 2023.

[71] V.V. Batyrev, On the stringy Hodge numbers of mirrors of quasi-smooth Calabi-Yau hypersurfaces, arXiv:2006.15825, 6.2020.

[72] M. Cranmer, Interpretable machine learning for science with PySR and SymbolicRegression.jl, arXiv:2305.01582, 2023.

[73] I. Biswas, G. Schumacher, Vector bundles on Sasakian manifolds, Adv. Theor. Math. Phys. 14 (2) (2010) 541–562.

[74] N. Tanaka, A Differential Geometric Study on Strongly Pseudo-Convex Manifolds, vol. 9, Kinokuniya, 1975, http://hdl.handle.net/2433/84914.

[75] T. King, S. Butcher, L. Zalewski, Apocrita - high performance computing cluster for Queen Mary University of London, https://doi.org/10.5281/zenodo.438045, Mar. 2017.