On Machine Learning Kreuzer-Skarke Calabi-Yau Manifolds

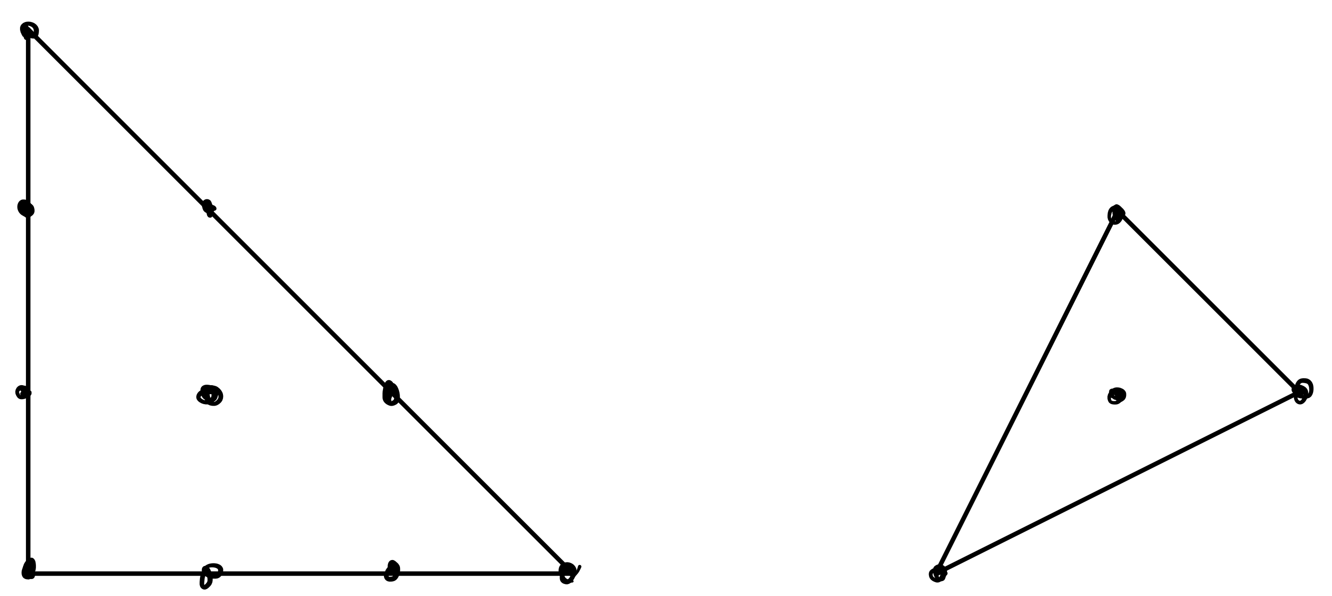
Per Berglund
University of New Hampshire

w/ Ben Campbell & Vishnu Jejjala, arXiv:2112:09117

string data 21, 12/17/21

Motivation & Results

- Use ML to learn topological data of Calabi-Yau threefolds from the Kreuzer-Skarke database of 473,800,776 reflexive polytopes [Kreuzer & Skarke]
- NN is able to learn/realize an exact expression for the Euler number in terms of minimum amount of input data from the polytope and its dual.
- For the individual Hodge numbers lower accuracy indicates lack of simple analytic expression.



Outline

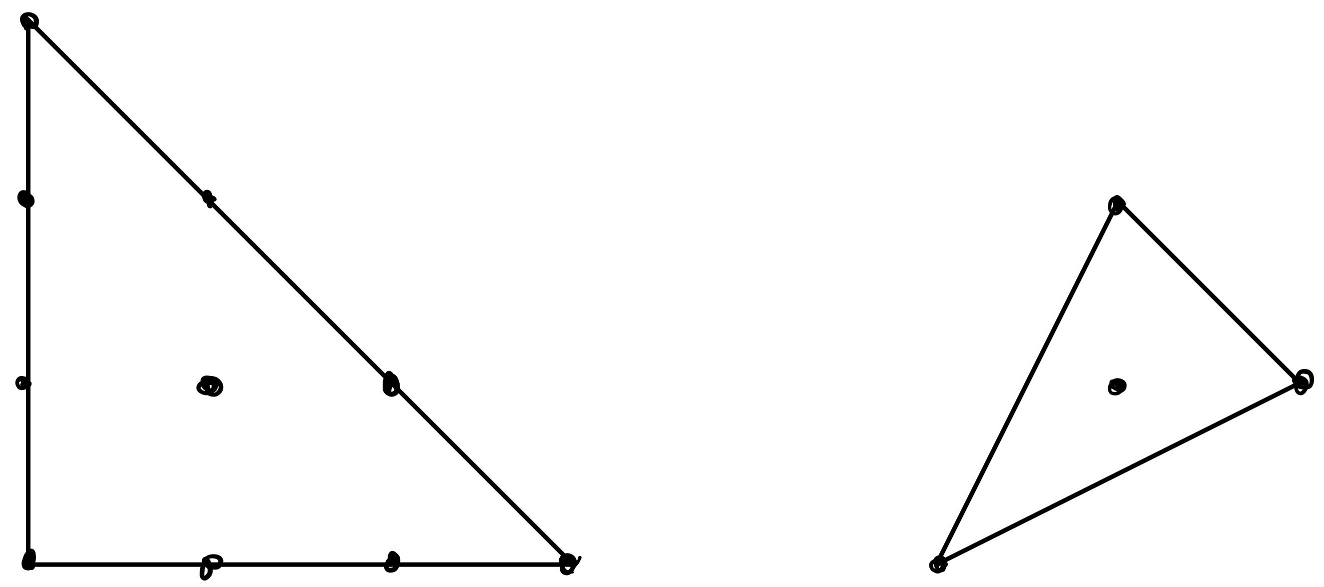
- Motivation and Results
- Background
- Reflexive Polytopes
- ML techniques
- Numerical Results
- Analytical Formulae
- Summary & Outlook

Background

- Large data sets of string vacua
  - Complete Intersection Calabi-Yau manifolds (CICY) [Candelas et al, Green & Hübsch]
  - Kreuzer-Skarke database of reflexive polytopes and Calabi-Yau hypersurfaces in toric varieties [Kreuzer & Skarke]
- ML has been successfully used in studying topological properties of CY manifolds, including obtaining exact analytical results [Brodie et al]
  - CICY in 3 dimensions [Bull et al, Erbin & Finetello]
  - CICY in 4 dimensions [He & Lukas]
  - CY hypersurfaces in weighted projective space [Berman et al]

Motivation & Results

- Use ML to learn topological data of Calabi-Yau threefolds from the Kreuzer-Skarke database of 473,800,776 reflexive polytopes
- NN is able to learn/realize an exact expression for the Euler number in terms of minimum amount of input data from the polytope and its dual.
- For the individual Hodge numbers lower accuracy indicates lack of simple analytic expression.



Kreuzer-Skarke & Reflexive Polytopes

- In n=2 dimensions, 16 reflexive polytopes
- In n=3 dimensions, 4319 reflexive polytopes [Kreuzer & Skarke]
- In n=4 dimensions, 473,800,776 reflexive polytopes [Kreuzer & Skarke]

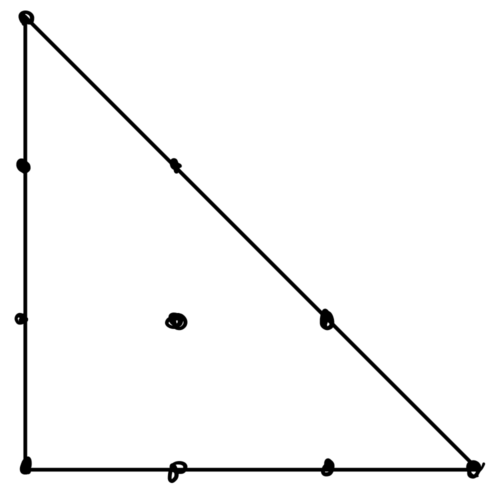
Reflexive polytope:
Convex hull of lattice polytope
with single interior point

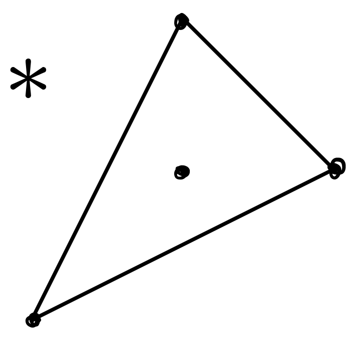
- Batyrev's mirror construction of CY M and W as hypersurfaces in toric varieties X_Δ and X_{Δ^*} with the ambient spaces constructed from given triangulation of the dual polytope Δ^* and Δ , respectively, [Batyrev, Batyrev & Borisov]

$$0 = \sum_{m\in\Delta} a_m \prod_i x_i^{\langle m, v_i^*\rangle+1} \qquad 0 = \sum_{m\in\Delta^*} b_m \prod_i y_i^{\langle m, v_i\rangle+1}$$

- The dual polytope is also reflexive and given by

$$\Delta^* = \{v \in \mathbb{R}^4 \mid \langle m, v\rangle \geq -1 \;\forall\, m \in \Delta\}$$

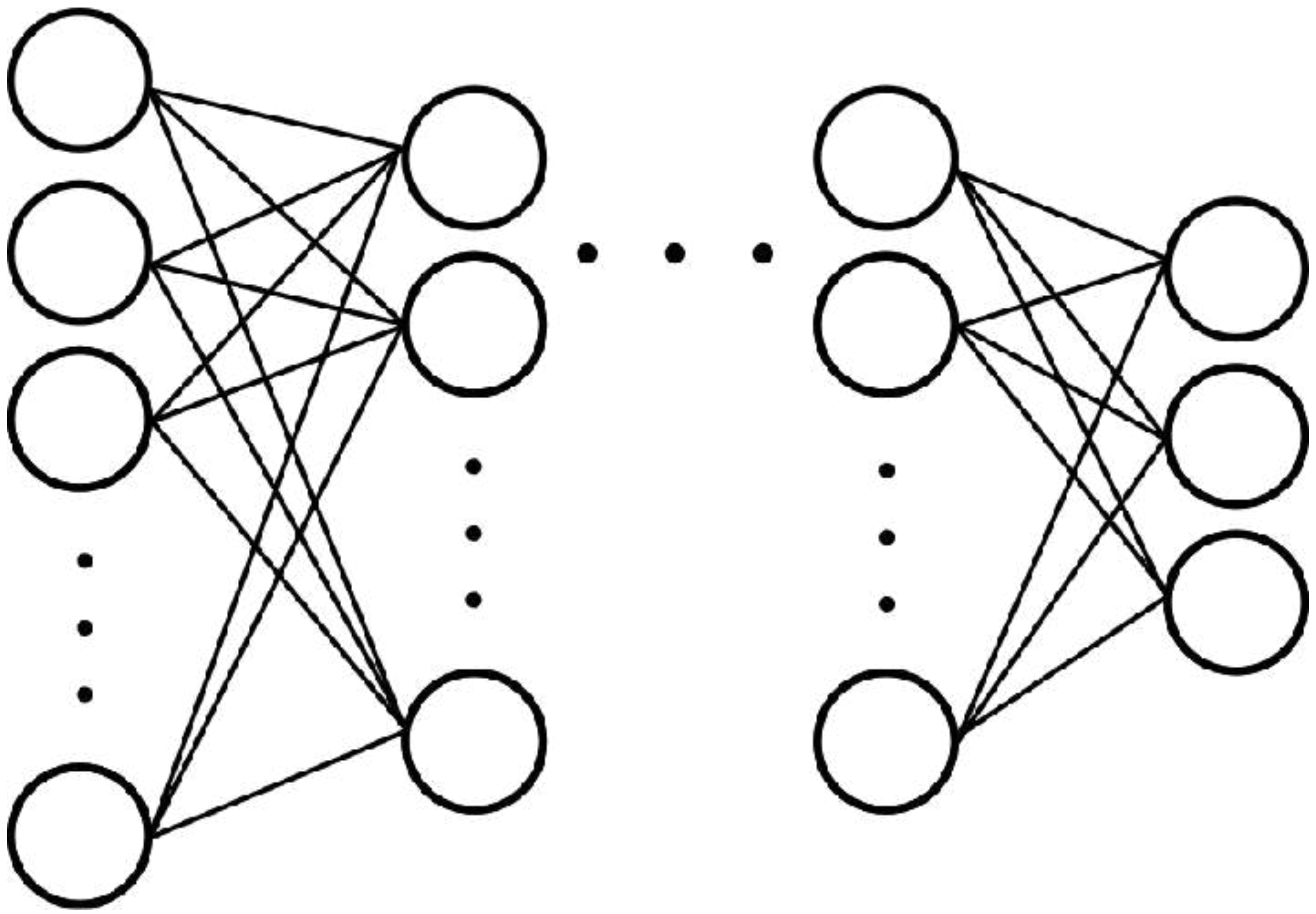
Δ 

Δ^* 

Neural Network

Implementation using Julia 1.6.2 and the Flux package

- Five hidden layers w/300 neurons/layer
- ReLu activation
- Adam algorithm + logit cross entropy loss fct
- 80/20 split of training/testing
  - 10^6 randomly selected 4d reflexive polytopes
  - 10^6 boundary 4d reflexive polytopes
- Input data $\boldsymbol{v} = (l(\Delta), l(\Delta^*), l(2\Delta), l(2\Delta^*))$
- Four different labels
  - Euler number
  - h11
  - h21
  - h11+h21



Numerical Results

Label	Accuracy (%)	Absolute Error	Relative Absolute Error
χ	97.36 ± 1.42	0.1746 ± 0.1941	0.0049 ± 0.0099
$h^{1,1}$	46.89 ± 0.91	0.7099 ± 0.0966	0.0222 ± 0.0029
$h^{2,1}$	46.74 ± 1.03	0.7262 ± 0.0896	0.0227 ± 0.0026
$h^{1,1} + h^{2,1}$	32.64 ± 0.27	1.464 ± 0.046	0.0214 ± 0.0006

Table 1. Mean accuracy, absolute error, and relative absolute error for model predictions on each label trained and tested on the randomly sampled data. Averages and standard deviations are taken over 100 models for each label. The total data is shuffled before being split into 80% training and 20% testing sets for each model.

Numerical Results

Label	Accuracy (%)	Absolute Error	Relative Absolute Error
χ	96.02 ± 1.24	0.2560 ± 0.0702	0.0035 ± 0.0018
$h^{1,1}$	75.35 ± 1.08	0.3142 ± 0.0494	0.0090 ± 0.0008
$h^{2,1}$	75.48 ± 0.78	0.3131 ± 0.0632	0.0089 ± 0.0009
$h^{1,1} + h^{2,1}$	67.77 ± 1.60	0.7122 ± 0.0649	0.0075 ± 0.0007

Table 2. Mean accuracy, absolute error, and relative absolute error for model predictions on each label trained and tested on the boundary data. Averages and standard deviations are taken over 100 models for each label. The total data is shuffled before being split into 80% training and 20% testing sets for each model.

Confusion Matrices

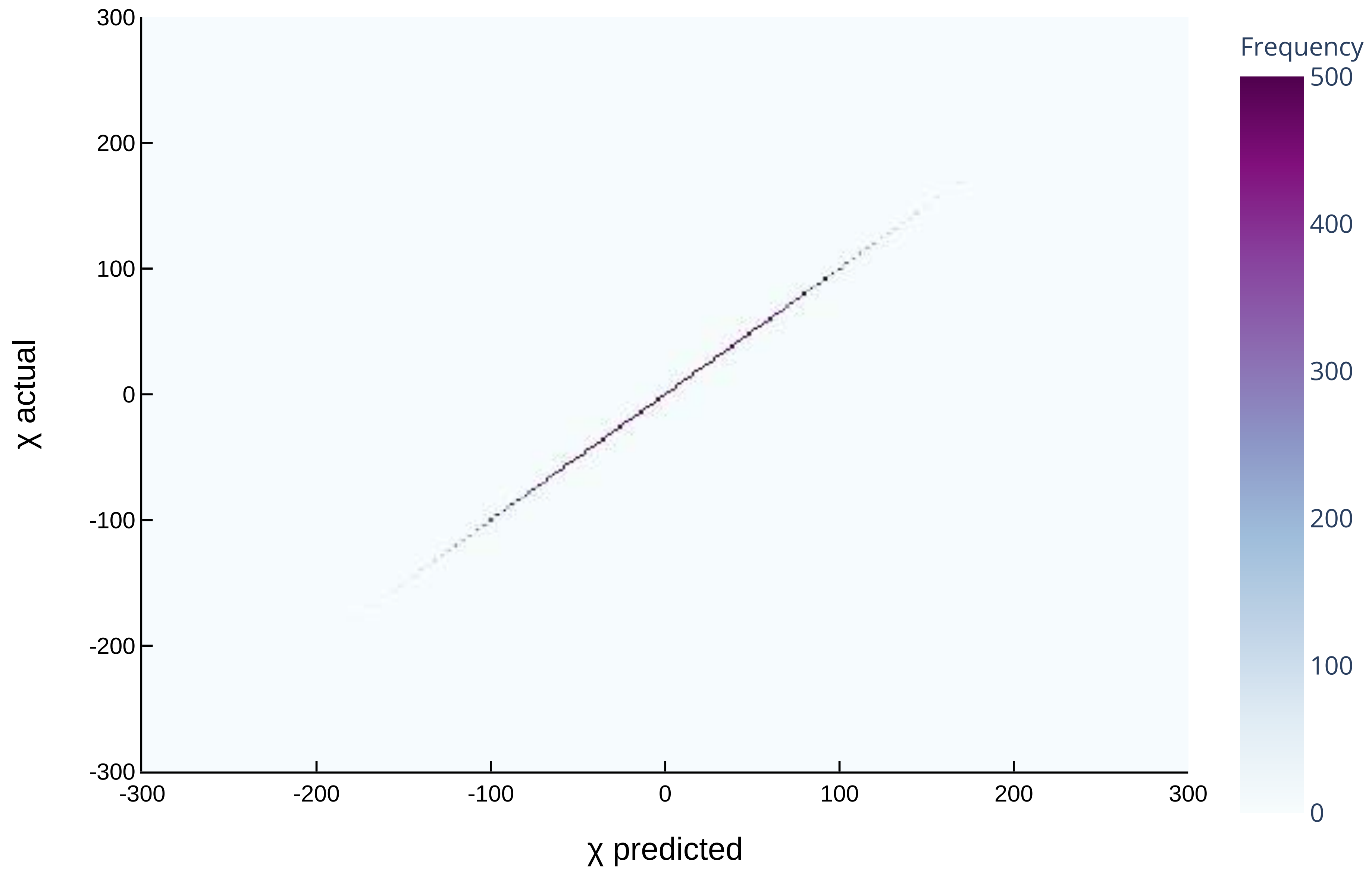


Figure 3. Confusion matrix for model trained on randomly sampled data evaluated on randomly sampled data. This model achieved an accuracy of 99.25% and mean absolute error of 0.0123. The range has been cropped to $\chi \in [-300, 300]$ as the majority of the data is in this interval.

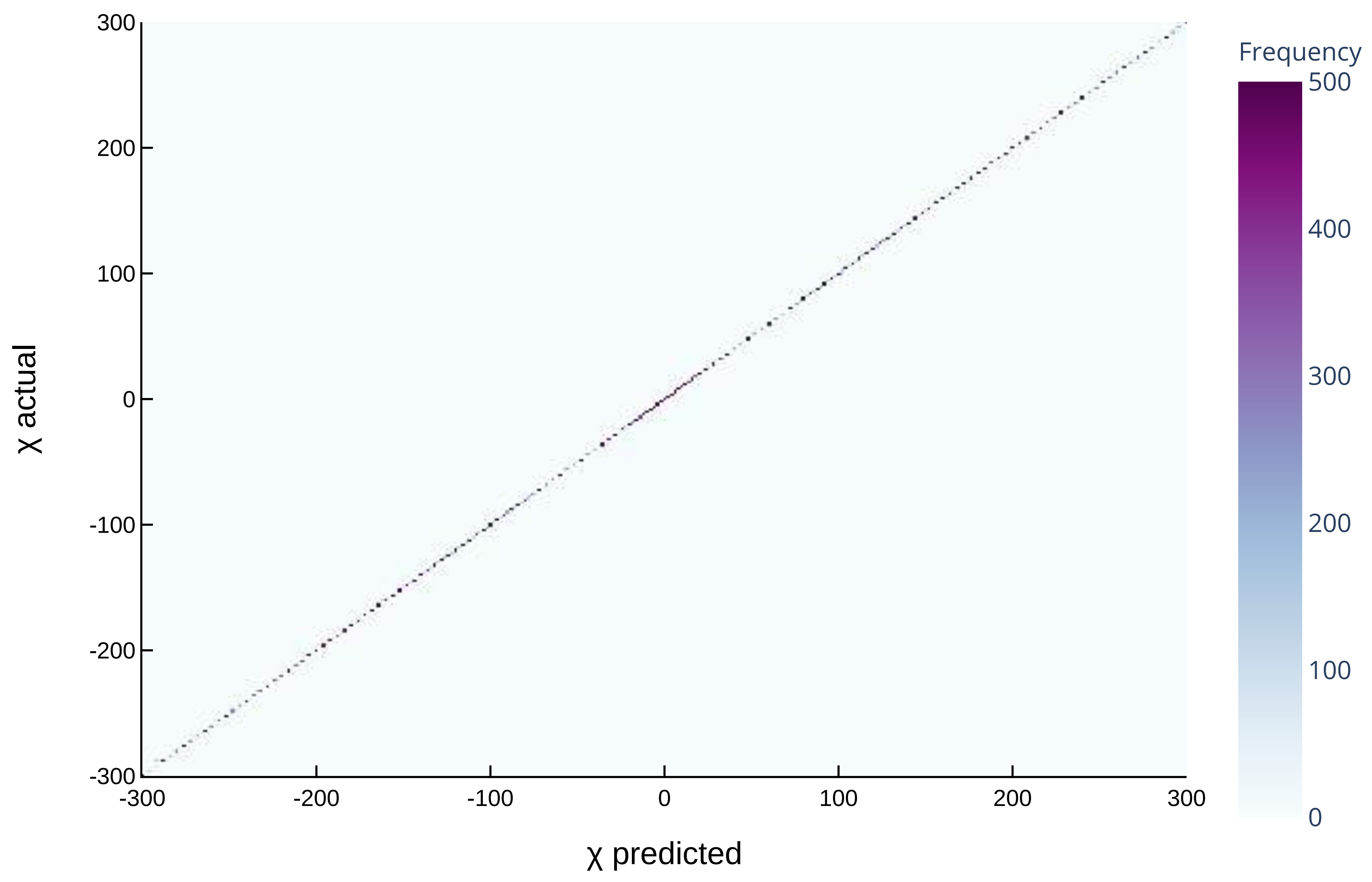


Figure 4. Confusion matrix for model trained on boundary data evaluated on boundary data. This model achieved an accuracy of 96.43% and a mean absolute error of 0.1339. The range has been cropped to $\chi \in [-300, 300]$ for comparison with Figure 3.

Linear Regression

Implementation using Mathematica 12.3.1

- High accuracy of ML predictions, especially for the Euler number, points to possible exact expression in terms of the input data
- Random sample of 200,000 reflexive polytopes and their duals.
- Applying linear regression on input data for 100,000 polytopes and their duals out of the above set, repeated 100 times:

$$\begin{aligned} h^{1,1} &= -(3.33 \pm 0.05) - (3.471 \pm 0.005) l(\Delta) + (5.529 \pm 0.005) l(\Delta^*) + \\ &\quad +(0.4176 \pm 0.0007) l(2\Delta) - (0.5824 \pm 0.0007) l(2\Delta^*) \ , \\ h^{2,1} &= -(3.33 \pm 0.05) + (5.529 \pm 0.005) l(\Delta) - (3.471 \pm 0.005) l(\Delta^*) + \\ &\quad -(0.5824 \pm 0.0007) l(2\Delta) + (0.4176 \pm 0.0007) l(2\Delta^*) \ . \end{aligned}$$

- Accuracy of 44% and 46%, respectively, when rounding to nearest integer, with an additional 44% and 42% allowing for prediction to be off by +/-1.

- Repeat for the sum and differences of the Hodge numbers

$$
\begin{aligned}
h^{1,1} + h^{2,1} \;&=\; -(6.64 \pm 0.01) + (2.06 \pm 0.01)\Big(l(\Delta) + l(\Delta^*)\Big) \\
&\quad -(0.165 \pm 0.001)\Big(l(2\Delta) + l(2\Delta^*)\Big)\ , \\
\frac{1}{2}\chi = h^{1,1} - h^{2,1} \;&=\; 9\Big(l(\Delta^*) - l(\Delta)\Big) + \Big(l(2\Delta) - l(\Delta^*)\Big)\ .
\end{aligned}
$$

- Lower accuracy for the sum of Hodge numbers, with correct accuracy only 23% and 38% when allowing for being off by +/- 1.
- Exact result for the Euler number!

Analytic Formulae

- Stringy Libgober-Wood identity, relating combinatorial data of polytope Δ and its dual Δ^* to topological data of the toric variety X_Δ [Batyrev & Schaller]

$$\sum_{i=0}^{n} \psi_i (i - \frac{n}{2})^2 = \frac{n}{12} d(\Delta) + \frac{1}{6} \sum_{\substack{\theta \in \Delta \\ \dim \theta = n-2}} d(\theta) d(\theta^*)$$

- Here the sum on the RHS is over the (n-2) faces with

$$d(\theta) = (n-2)! \mathrm{Vol}(\theta)$$

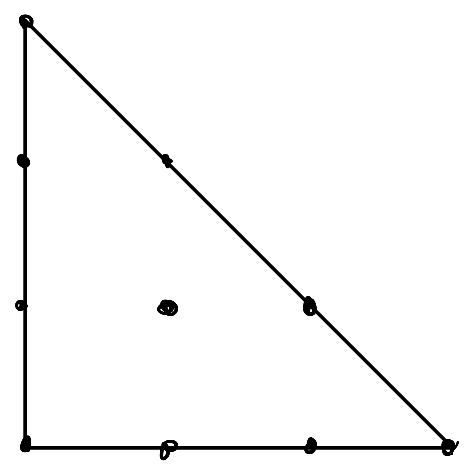
- $\psi_i(\Delta)$ encodes topological information about X_Δ

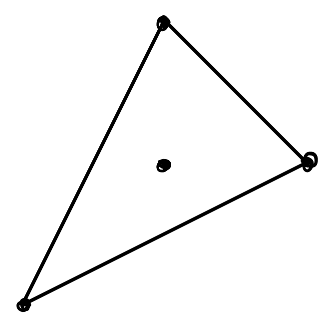
- Ehrhart series/polynomial, with $k\Delta$ the kth scaled-up polytope [Ehrhart, Danilov]

$$P_\Delta(t) = \sum_{k\geq 0} l(k\Delta)t^k \qquad P_\Delta(t) = \frac{\Phi(t)}{(1-t)^{n+1}} \qquad \Phi(t) = \sum_{i=0}^{n} \psi_i(\Delta)t^i$$

- For n=2 this is the 12-theorem:

$$12 \quad = \quad d(\Delta) \ + \ d(\Delta^*)$$





- For n=3 this implies the so called "24-theorem" or simply that

$$\chi(M) = \int_X c_1(X)c_2(X) \ = \sum_{\substack{\theta\in\Delta \\ \dim\theta=1}} d(\theta)d(\theta^*) = 24$$

- For n=4 we get for the polytope

$$12\,(l(\Delta)-1) = 2d(\Delta) + \sum_{\substack{\theta\in\Delta \\ \dim\theta=2}} d(\theta)d(\theta^*)$$

- Similarly for the dual polytope

$$12\,(l(\Delta^*)-1) = 2d(\Delta^*) + \sum_{\substack{\theta^*\in\Delta^* \\ \dim\theta^*=2}} d(\theta^*)d(\theta)$$

- The Euler number can be written [Batyrev]

$$\chi(M) = \int_X \left[c_1(X)c_3(X) - c_1^2(X)c_2(X)\right] = \sum_{\substack{\theta\in\Delta \\ \dim\theta=1}} d(\theta)d(\theta^*) - \sum_{\substack{\theta\in\Delta \\ \dim\theta=2}} d(\theta)d(\theta^*)$$

- Thus,

$$\chi(M) = 12\,(l(\Delta^*) - l(\Delta)) + 2\,(d(\Delta) - d(\Delta^*))$$

- Finally, we use that

$$d(\Delta) = 2 + l(2\Delta) - 3l(\Delta)\ , \qquad d(\Delta^*) = 2 + l(2\Delta^*) - 3l(\Delta^*)$$

- Thus

$$\chi(M) = 2\,(l(2\Delta) - l(2\Delta^*)) + 18\,(l(\Delta^*) - l(\Delta))$$

- Example: quintic hypersurface in P^4

$$l(\Delta) = 126,\ l(\Delta^*) = 6 \quad l(2\Delta) = 1001\; l(2\Delta^*) = 21$$

- This gives

$$\chi(M) = -200$$

Summary

- Analyzed a large (randomly selected) set of reflexive polytopes from the n=4 Kreuzer-Skarke database using ML.
- NN is able to learn/extract/discover an analytic expression for the Euler number given a very limited input data: $\boldsymbol{v} = (l(\Delta), l(\Delta^*), l(2\Delta), l(2\Delta^*))$

$$\chi(M) = 2\,(l(2\Delta) - l(2\Delta^*)) + 18\,(l(\Delta^*) - l(\Delta))$$

- The accuracy for predicting the individual Hodge numbers varies from 46% to 75%, indicating that such a simple expression does not exist for, except for the so called favorable cases where

$$h^{1,1} = l(\Delta^*) - 5. \qquad \text{and/or} \qquad h^{2,1} = l(\Delta) - 5$$

Outlook

- What is required (additional input data) to learn other topological data?
- How to extend analysis to n=5—can we use ML to to study elliptically fibered CY fourfold?
  - No complete classification of reflexive polytopes exists—can ML be used in this classification?
- More general polytopes/generalized constructions of CYs—can ML be extended beyond reflexive polytopes
  - gCICY
  - VEX polytopes/triangulations