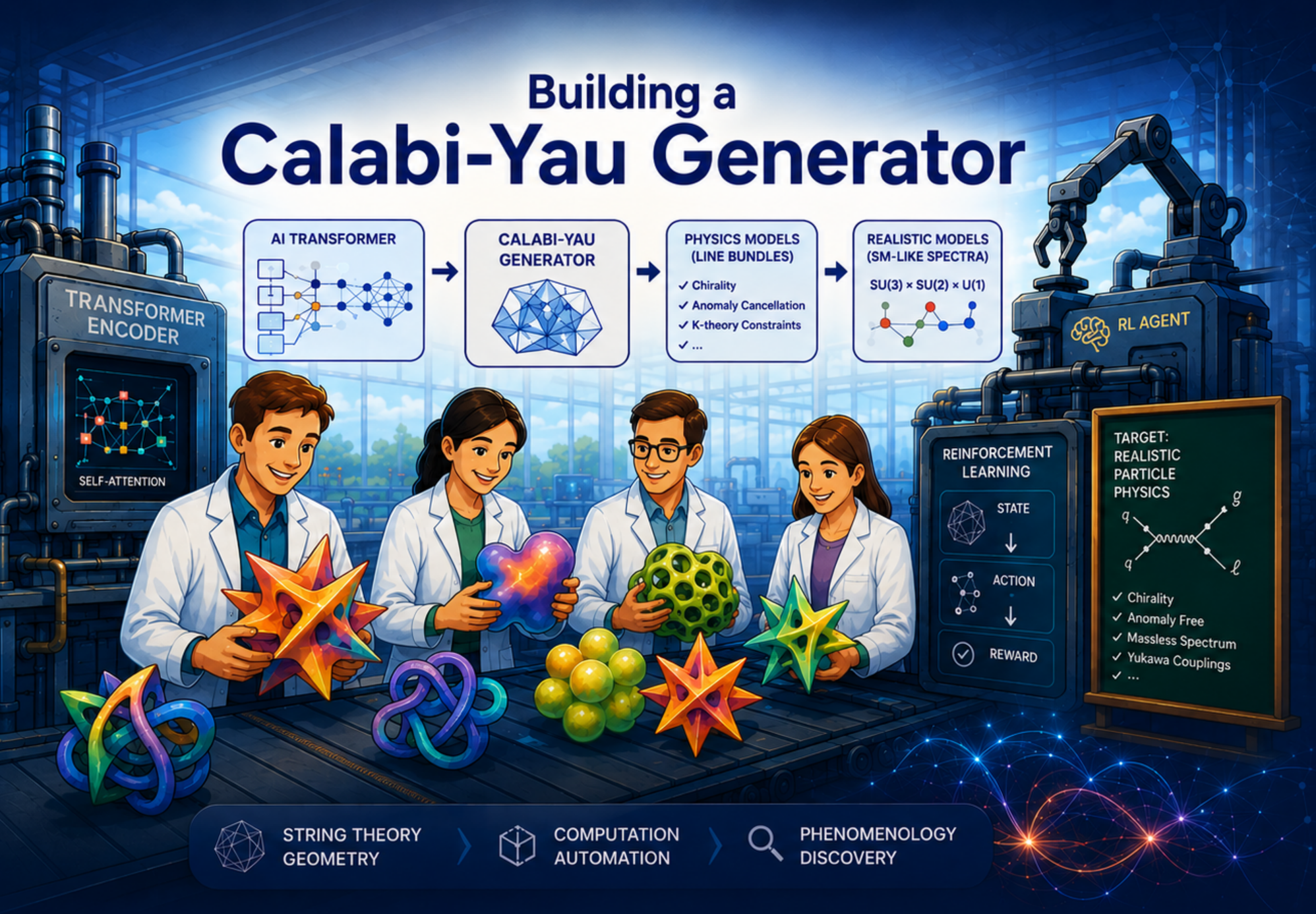
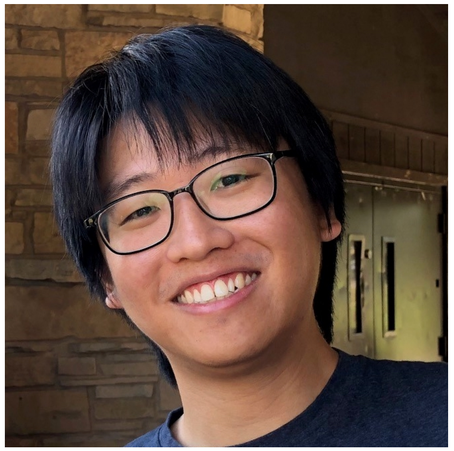
Building a Calabi-Yau Generator



Gary Shiu

University of Wisconsin–Madison

The Cast



Jacky Yip

Charles Arnal

Francois Charton

Alessandro Mininno

- Jacky H.T. Yip, Charles Arnal, Francois Charton, and Gary Shiu, *"Transforming Calabi-Yau Constructions: Generating New Calabi-Yau Manifolds with Transformers,"* arXiv:2507.03732, to appear in ATMP.
- Jacky H.T. Yip, Alessandro Mininno, and Gary Shiu, "Exploring Heterotic Standard Models with Transformers, arXiv: 2605.xxxxx.

The String Landscape

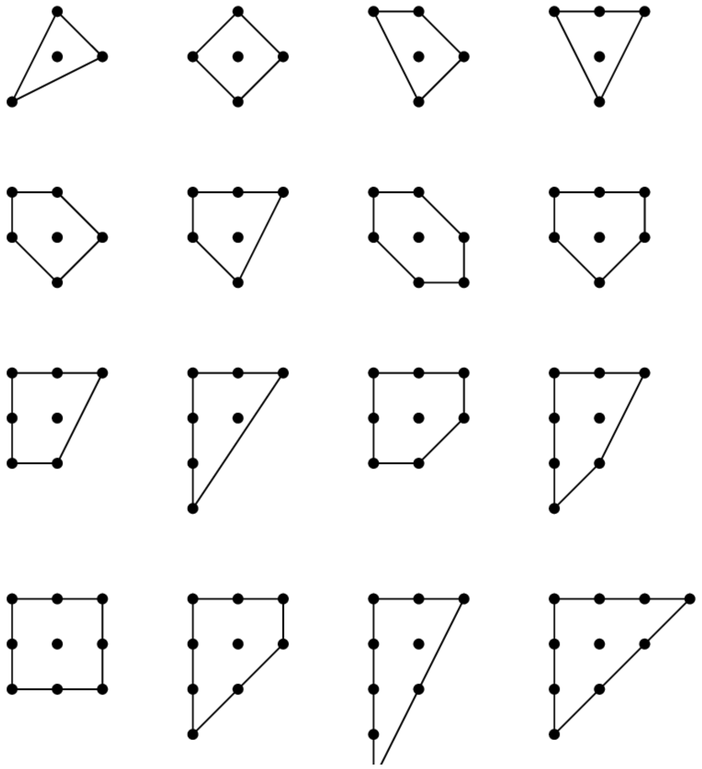
- The vastness of the string landscape presents a serious computational challenge.
- The immensity stems from the multitude of choices of
  - compactification manifold (or for non-geometric constructions, choice of CFT)
  - bundle and/or brane configurations
  - quantized fluxes
  - …
- Yet, the # of string vacua (with a given cutoff on compactification volume) is conjectured to be **finite**
  - An underlying premise in the program of landscape statistics [Douglas, ’05];[Acharya, Douglas, ’06]
  - Universal properties of quantum gravity [Vafa, ’05];[Hamada, Montero, Vafa, Valenzuela, ’21]

Finiteness and Enumeration

- However, only when we restrict to very small regions of the landscape that an exact number of solutions is known, e.g.,
  - 134,474,650,261 intersecting brane models in a *particular* orientifold [Loges, GS, '22] (though it was proven earlier that the number is finite [Douglas, Taylor, '06])
- Calabi-Yau manifolds are a well-motivated class of solutions for data mining the landscape:
  - Yau conjectured there are finitely many topological types of CY manifolds in each dimension.
  - CY 3-folds (or 4-folds for F-theory) yield 4d vacuum configurations that can accommodate realistic particle physics (see [Marchesano, GS, Weigand, '24] for recent review]).
  - While CY 3-folds are neither fully classified nor known to be finite in number, those that can be realized as hypersurfaces in toric varieties are amenable to combinatorial enumeration.

Toric Calabi-Yau Constructions

- Batyrev's construction: each `fine regular star triangulation' (FRST) of a **4d reflexive polytope** $\Delta \subset \mathbb{Z}^4$ yields a toric variety whose generic anticanonical divisor is a smooth Calabi-Yau.



all 16 reflexive polytopes in 2d, up to lattice automorphism

- 4d reflexive polytopes have been fully enumerated (473,800,776 of them) [Kreuzer, Skarke, '00].
- The formidable combinatorics lie in the FRSTs whose number grows exponentially with N_{vert} of the polytope.
- An FRST is a triangulation $\mathscr{T}$ of Δ such that it is:
  - Fine: every point is a vertex for some simplex in $\mathscr{T}$
  - Regular: $\mathscr{T}$ is a projection from one higher dim.
  - Star: all full dim simplices in $\mathscr{T}$ have origin as vertex.

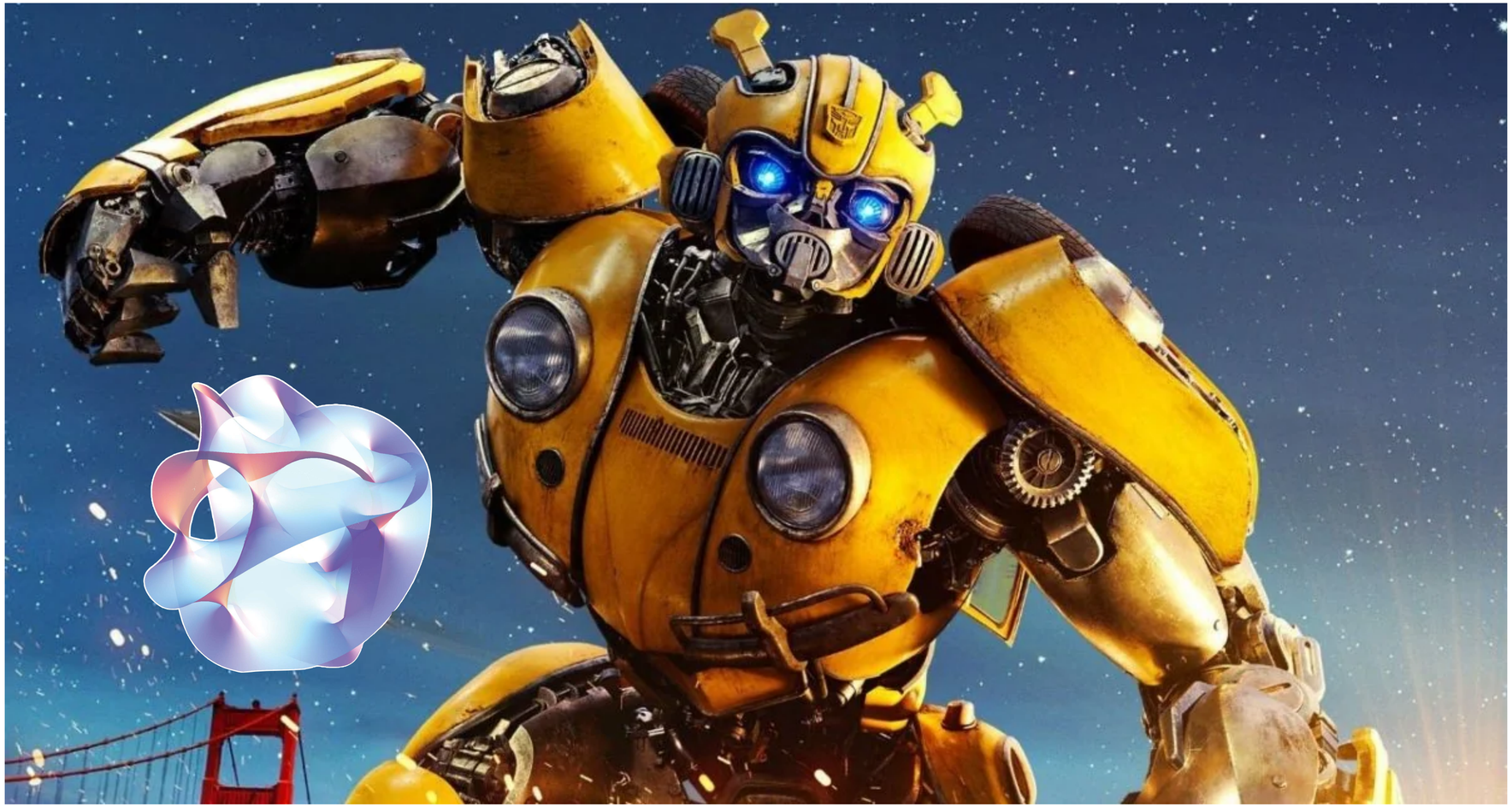
Learning Toric Calabi-Yau

- On the other hand, different FRSTs can give rise to topologically equivalent CY 3-folds.
- Enumeration is unfeasible, though $N_{CY} < 1.2 \times 10^{296}$ [McFadden, Orevkov, Stepniczka, '26].
- Software packages such as CYTools have sped up the triangulation of polytopes, but it would take longer than the age of the universe to uncover a sizable fraction of toric CY 3-folds!
- This calls for a scalable, self-improving learning algorithm to automate the generation of CYs:
  - Non-learning algorithms do not scale well with polytope size. Moreover, no lessons learned.
  - Genetic algorithms [MacFadden, Schachner, Sheridan, '24] and reinforcement learning [Berglund, Butbaia, He, Heyes, Hirst, Jejjala, '24] were used to perform targeted searches for CYs with favorable properties, not in expanding the search space. Moreover, it is not clear if lessons are transferable to other polytopes.
  - We develop a transformer model to automate CY generation [Yip, Arnal, Charton, GS, '25].

CYTranformer

- We demonstrate success of our transformer [Yip, Arnal, Charton, GS, ’25] in finding **new** CYs:
  - Training on polytopes with N_{vert} , our model learns to predict FRSTs for new polytopes with $N \geq N_{\text{vert}}$.
  - The FRSTs generated are representative of the ensemble, as quantified by several measures.
- We further show that our ML model can automate the generation of new toric CYs in a **self-improved** manner:
  - Training on a relatively small number of simple polytopes, the model learns to bootstrap its way up to more complex polytopes. Self-improvement can be further enhanced with **priming**.
  - The transfer of knowledge is indicative that the ML model learns the distribution of FRSTs in the space of all triangulations of reflexive polytopes.

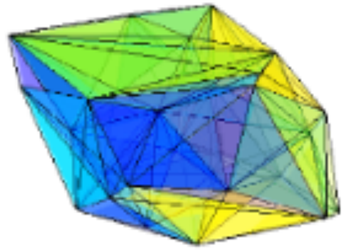
Transforming Calabi-Yau Constructions



GARY SHIU, UNIVERSITY OF WISCONSIN-MADISON

Learning to Triangulate Polytopes

- A triangulation of a reflexive polytope is represented by a sequence of simplices.

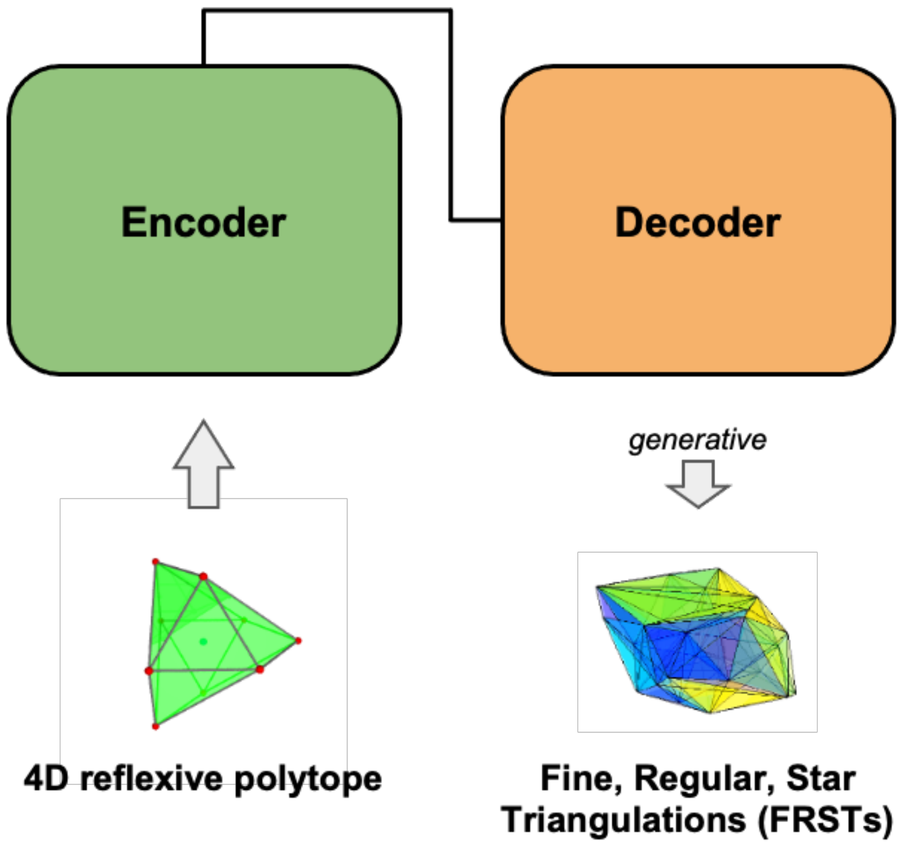


$$\mathscr{T} = \{\{\mathbf{0}, V_1, V_2, V_3, V_4\}, \{\mathbf{0}, V_1', V_2', V_3', V_4'\}, \dots\}$$

- An FRST is a triangulation satisfying a specific syntax → naturally a language problem.
- Suppose we use a **hypergraph neural network** to learn FRSTs: each node represents a polytope vertex; hyperedge features are learned probabilities of the corresponding simplices being in an FRST.
- What is learned from the hypergraph NN is the likelihood of a 4-simplex being in an FRST of the given polytope **on average**, which provides little information on distinct individual FRSTs.

Triangulating Polytopes with Encoder Decoder Transformer

[Yip, Arnal, Charton, GS, ’25]



- Transformers [Vaswani et al, ‘17] are designed for sequence modeling tasks e.g. natural language processing.
- Key to transformers is the **attention mechanism**.
- Token-by-token generative framework enables transformers to learn the distribution of data in an autoregressive manner.
- In our model, each token represents a simplex: the probability of generating the next simplex is **conditioned** on all previously generated ones.
- The sequence of simplices generated by our transformer is a candidate FRST.

Embedding

- A 4d reflexive polytope is an $(N_{\text{vert}} - 1) \times 4$ array of integer-valued coordinates of the vertices.
- An example for $N_{\text{vert}} = 10$:

```
[[-1  0  0  0]
 [-1  2  4 -1]
 [-1  1 -1  1]
 [-1  1  2  0]
 [ 1 -1 -1  0]
 [-1  0 -1  1]
 [ 0  0 -1  0]
 [-1  0 -1  0]
 [-1  1 -1  0]]
```

- The encoder embedding layer learns to convert coordinates of a point in 4d to a vector in the embedding space.

Tokenization

- The decoder input is a sequence of 4-simplices: each 4-simplex is a token.
- The token representing a simplex depends on the order of the vertices in the encoder input.
- For example, $\langle 0 \rangle$ represents the 4-simplex composed of $\{\text{origin}, V_0, V_1, V_2, V_3\}$ and $\langle 99 \rangle$ represents the 4-simplex composed of $\{\text{origin}, V_2, V_3, V_6, V_8\}$.
- **Location, location, location:** the position embedding layer in the encoder is crucial to our tokenization.
- Permuting the vertices does not change the polytopes. With enough training, the transformer can learn the symmetries.
- As we scale up the project, we may employ a permutation-invariant architecture e.g., a set transformer [Lee et al, '19]., or to use a tokenization that does not encode the ordering of the simplifies in the triangulation.

Vocabulary

- The vocabulary consists of the following tokens:

$$\underbrace{\{\mathbf{0}, V_1, V_2, V_3, V_4\}, \{\mathbf{0}, V_1', V_2', V_3', V_4'\}, \dots}_{\binom{N_{\text{vert}}-1}{4} \text{ distinct 4-simplices}}, \quad \underbrace{\langle \text{sos} \rangle}_{\text{start of sentence}}, \quad \underbrace{\langle \text{eos} \rangle}_{\text{end of sentence}}, \quad \underbrace{\langle \text{pad} \rangle}_{\text{padding}}$$

- An example of a decoding input:

```
[<sos> <20> <32> <90> <108> <36> <121> <47> <62> <2> <54> <125> <69> <84>
<101> <91> <56> <6> <5> <57> <13> <7> <eos> <pad> <pad> <pad> <pad> <pad>
<pad> <pad>].
```

- The size of the vocabulary suffers from combinatorial explosion (as $N_{\text{vert}} \sim \mathcal{O}(10^2)$). This can be mitigated by using multiple tokens to represent each 4-simplex, reducing vocabulary size.

Regularity and Height Vectors

- A regular triangulation can be constructed by lifting the vertices of a polytope into one higher dimension and projecting the lower codimension-1 faces of the convex hull.

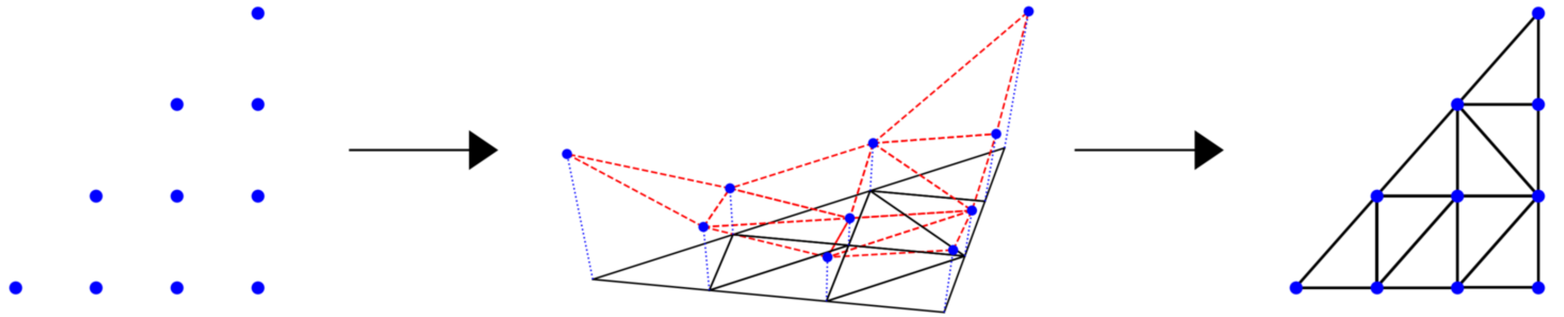
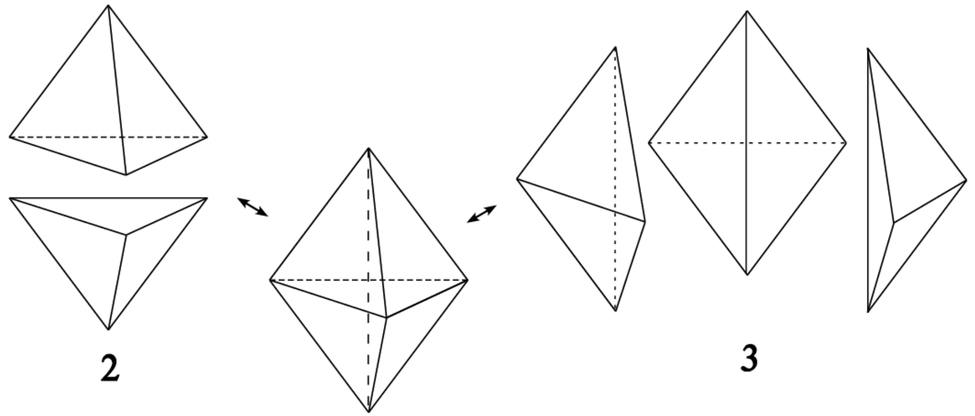


Figure from 2008.01730

- If a lifted point lies in the interior of the convex hull, then the resulting triangulation does not include that vertex.
- Not all projections result in triangulations, but rather in regular subdivisions that do not fully divide the polytopes into simplices.
- Fineness is easy to check, and the star requirement is met by sufficiently lowering the origin.

Delauney Triangulation and Sampling Algorithms

- For small polytopes ($h_{1,1} \leq 7$), all FRSTs can be enumerated by TOPCOM (exploring bistellar flips).
- The Delauney triangulation can be constructed by
  - Choosing $h_i = |\mathbf{p}_i|^2$, where $\mathbf{p}_i$ = position vector of vertices
  - Subdividing further into simplices makes it fine.
  - Lower the origin makes it star.

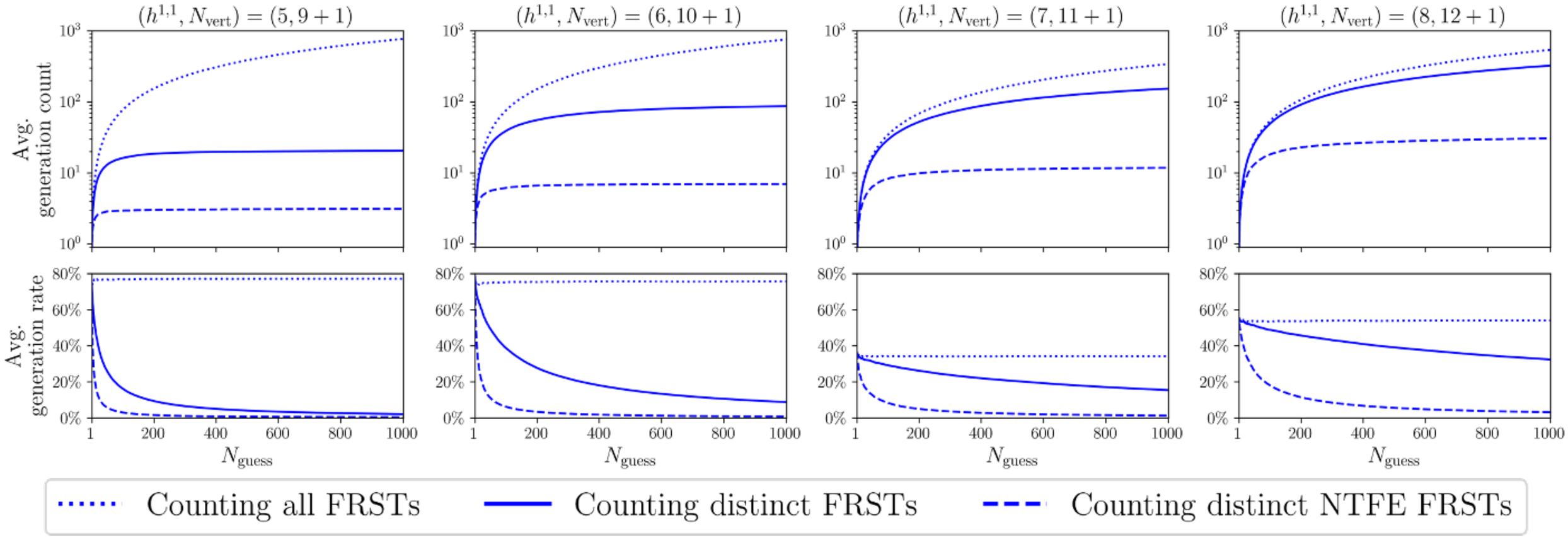


- Fast algorithm [Demirtas, McAllister, Rios-Tascon, '22]:
  - Initiate the Delaunay triangulation.
  - Sample ϵ_i for each point i from a Gaussian distribution with a standard deviation σ , and update the height vectors by setting $h_i \to h_i + \epsilon_i$.
  - Check if the resulting triangulation is fine and, if not, repeat sampling for ϵ_i and updating h_i.
- The fast algorithm can be combined with random flips to give more representative FRSTs, but this fairer sampling is exceedingly slow since the number of flips grows super-exponentially with N_{vert}.

Wall's Theorem and Two Face Equivalence

- Wall's theorem (1966) implies that simply connected Calabi-Yau threefolds with torsion-free homology are completely classified by the Hodge numbers, triple intersection numbers and second Chern class.
- Two FRSTs whose restrictions to the 2-faces of a polytope are identical define homotopy equivalent CYs.
  - The Hodge numbers depend only by the polytope and not its triangulation.
  - The triple intersection numbers and the second Chern class are specified by the restrictions of the triangulation to the 2-faces of the polytope [Demirtas, McAllister, Rios-Tascon, '20].
- A measure of representativeness of our algorithm is its ability to generate topologically distinct CYs, or non-two-face-equivalent (NTFE) FRSTs.

CYTransformer: FRST Generation Count and Rate

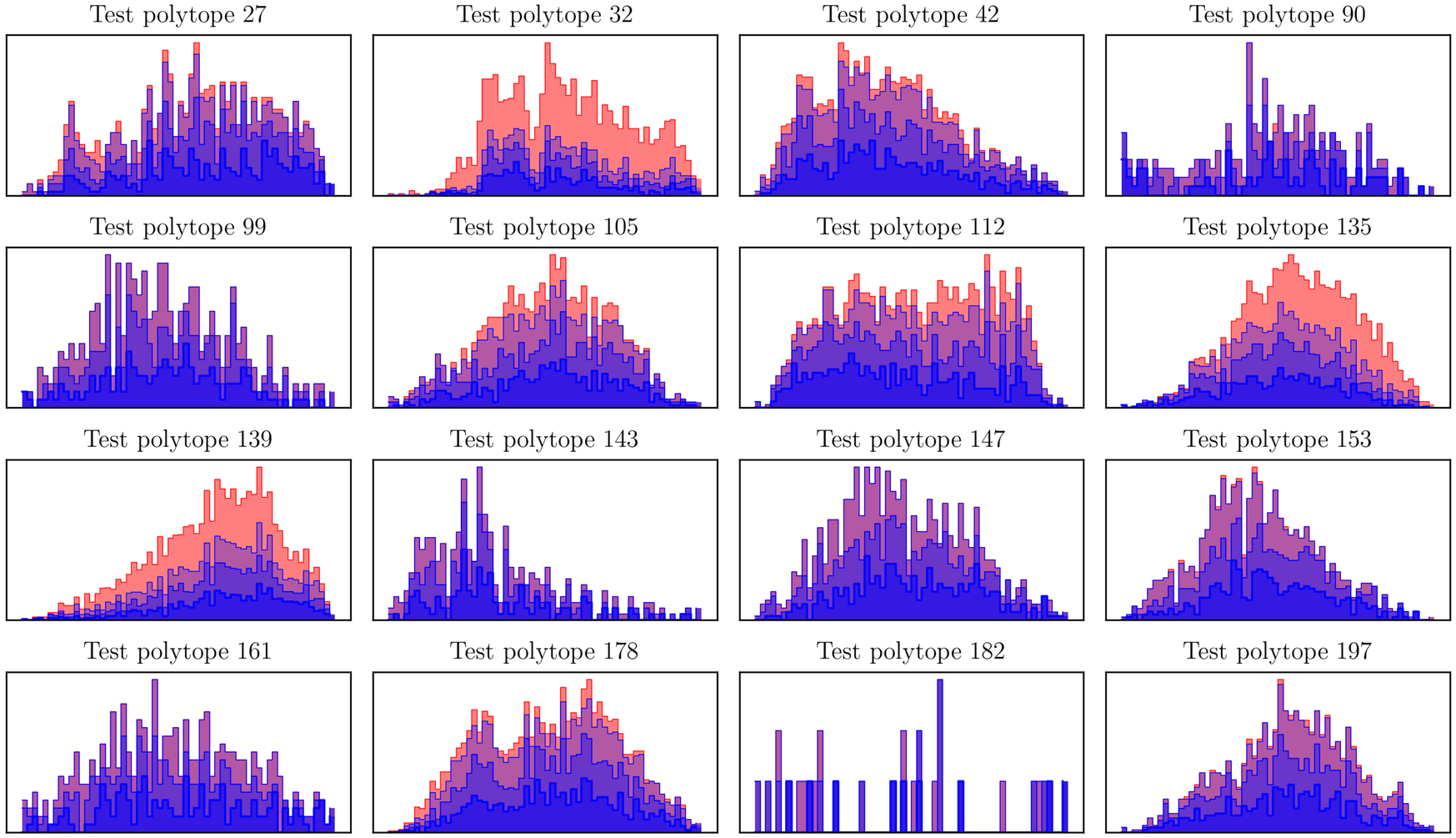


- Cumulative count per polytope averaged over 200 test polytopes (up to 1000 inference runs).
- Small polytopes: FRST space quickly exhausted; large polytopes: continued exploration.

Representativeness

- In principle, we can use flip distance, which measures the minimal number of bistellar flips needed to transform one triangulation into another, as a proxy for closeness of triangulations.
- However, the number of possible flips grows super-exponentially with the number of vertices. Computing the shortest paths in the flip graph is intractable even with just a few vertices.
- We use height vector similarity as a proxy of how widely the model explores the space of FRSTs.
- There is a subtlety: height vectors differ by an affine-linear function $\mathbf{h}_i = \mathbf{h}_i + c_0\mathbf{1} + c_j\mathbf{p}_i^j$ as well as overall rescaling of the height vectors give the same triangulation.
- We remove this ambiguity by projecting each height vector orthogonally to the affine subspace.
- We compute cosine similarity between each projected height vector and that of the unique Delaunay triangulation which is insensitive to the overall rescaling.

CYTransformer: Height Vector Distribution



$$(h^{1,1}, N_{\text{vert}}) = (8, 12+1)$$

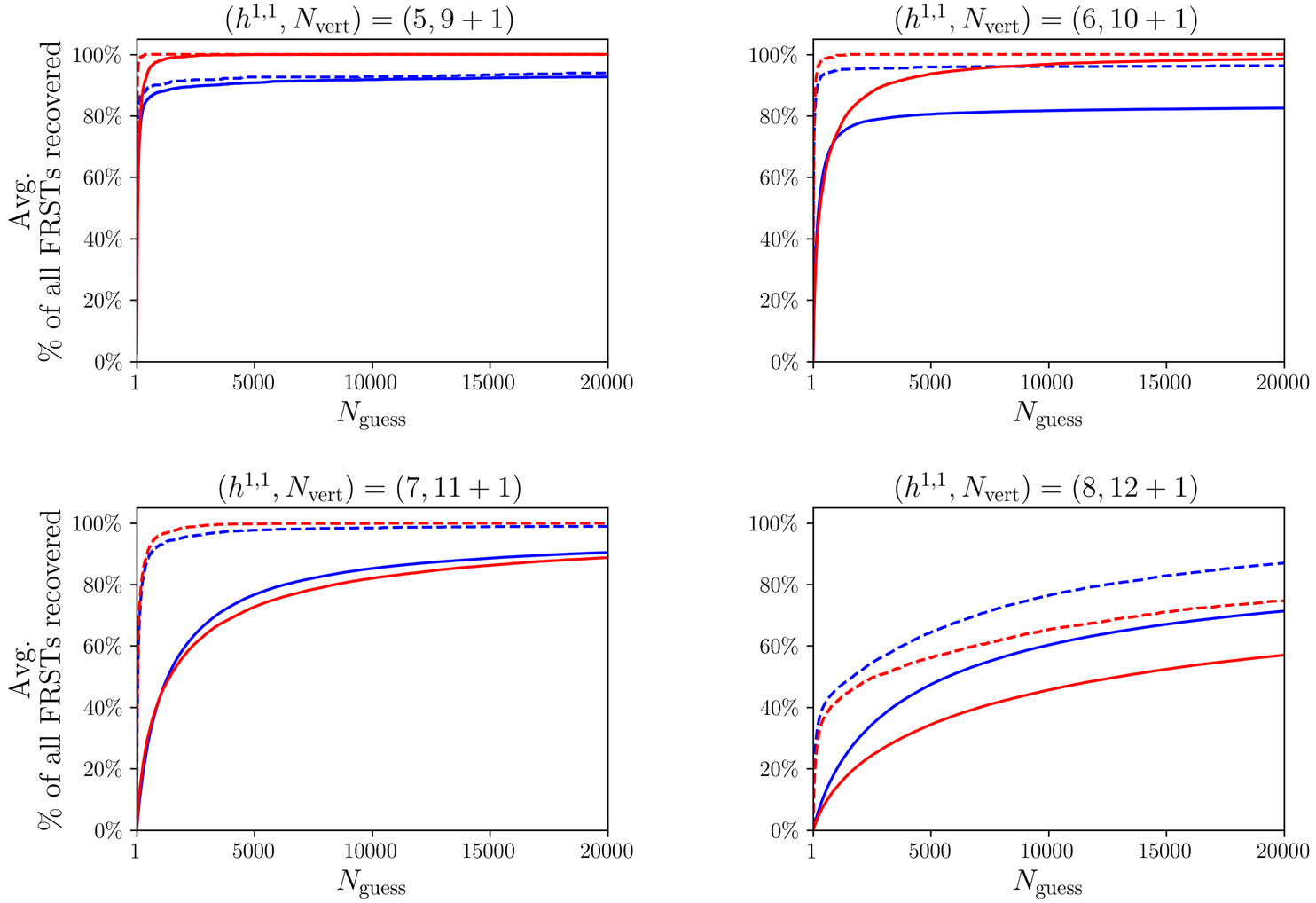
CYTransformer - 33% of distinct FRSTs recovered
CYTransformer - 67% of distinct FRSTs recovered
CYTransformer - 100% of distinct FRSTs recovered
Complete set of distinct FRSTs of the test polytope

$$h^{1,1}(X) = \ell(\Delta^\circ) - 4 - 1 - \sum_{\Gamma^\circ} \ell^*(\Gamma^\circ) + \sum_{\Theta^\circ} \ell^*(\Theta^\circ)\ell^*(\hat{\Theta}^\circ)$$

$$h^{2,1}(X) = \ell(\Delta) - 4 - 1 - \sum_{\Gamma} \ell^*(\Gamma) + \sum_{\Theta} \ell^*(\Theta)\ell^*(\hat{\Theta}),$$

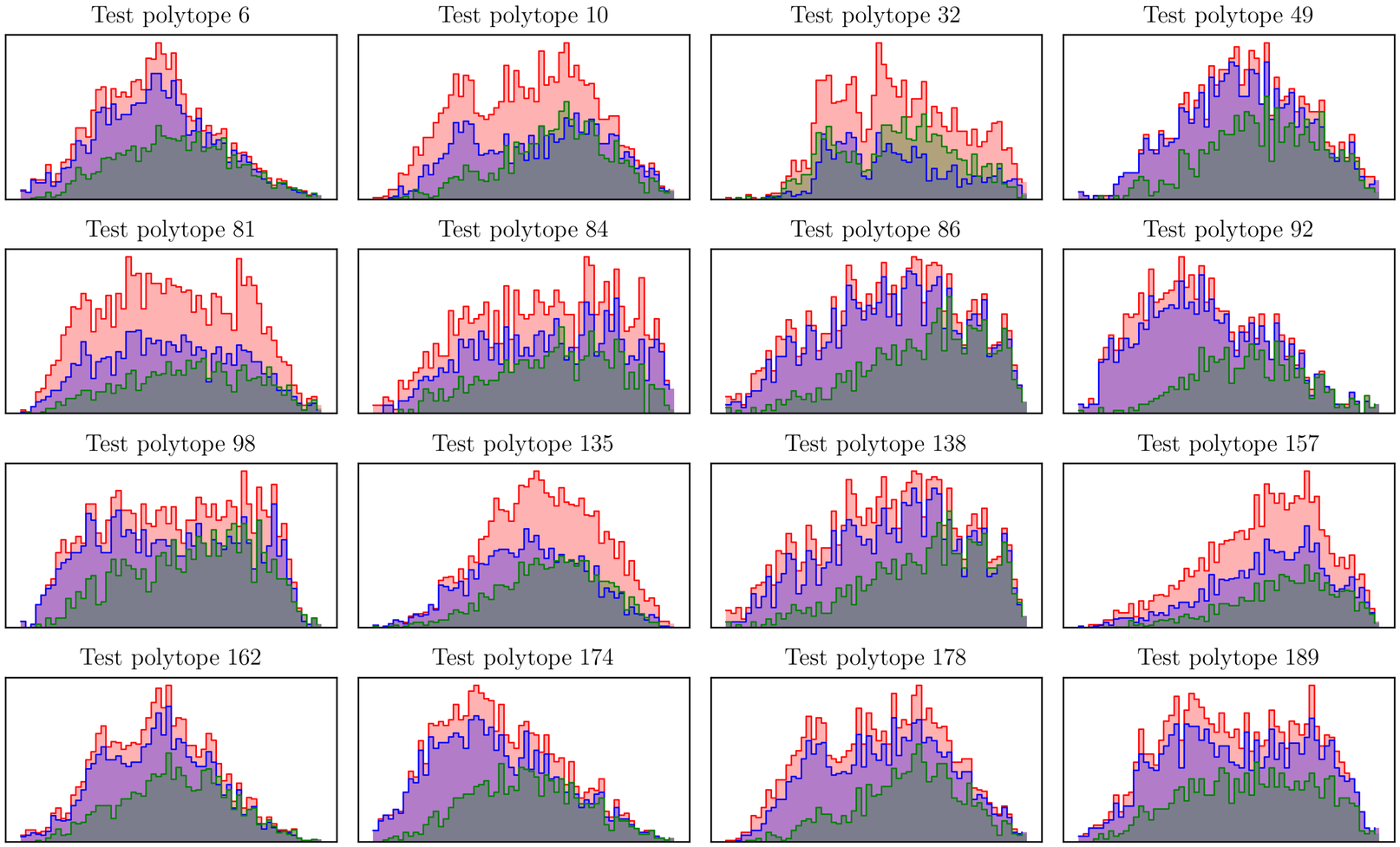
GARY SHIU, UNIVERSITY OF WISCONSIN-MADISON

Comparison: FRST Recovery Curve



- Averaged over 200 test polytopes
- For **small** polytopes, **fast sampler** performs well
  - Fast sampler scans small FRST space efficiently and thoroughly
- For **large** polytopes, **CYTransformer** clearly outperforms
  - CYTransformer explores the FRST space unbiasedly
- (7, 11+1) is where machine learning beats brute-force randomness
- **Fast sampler is a local scanner; CYTransformer is a global explorer**

Comparison: Height Vector Distribution



$$(h^{1,1}, N_{\text{vert}}) = (8, 12 + 1)$$

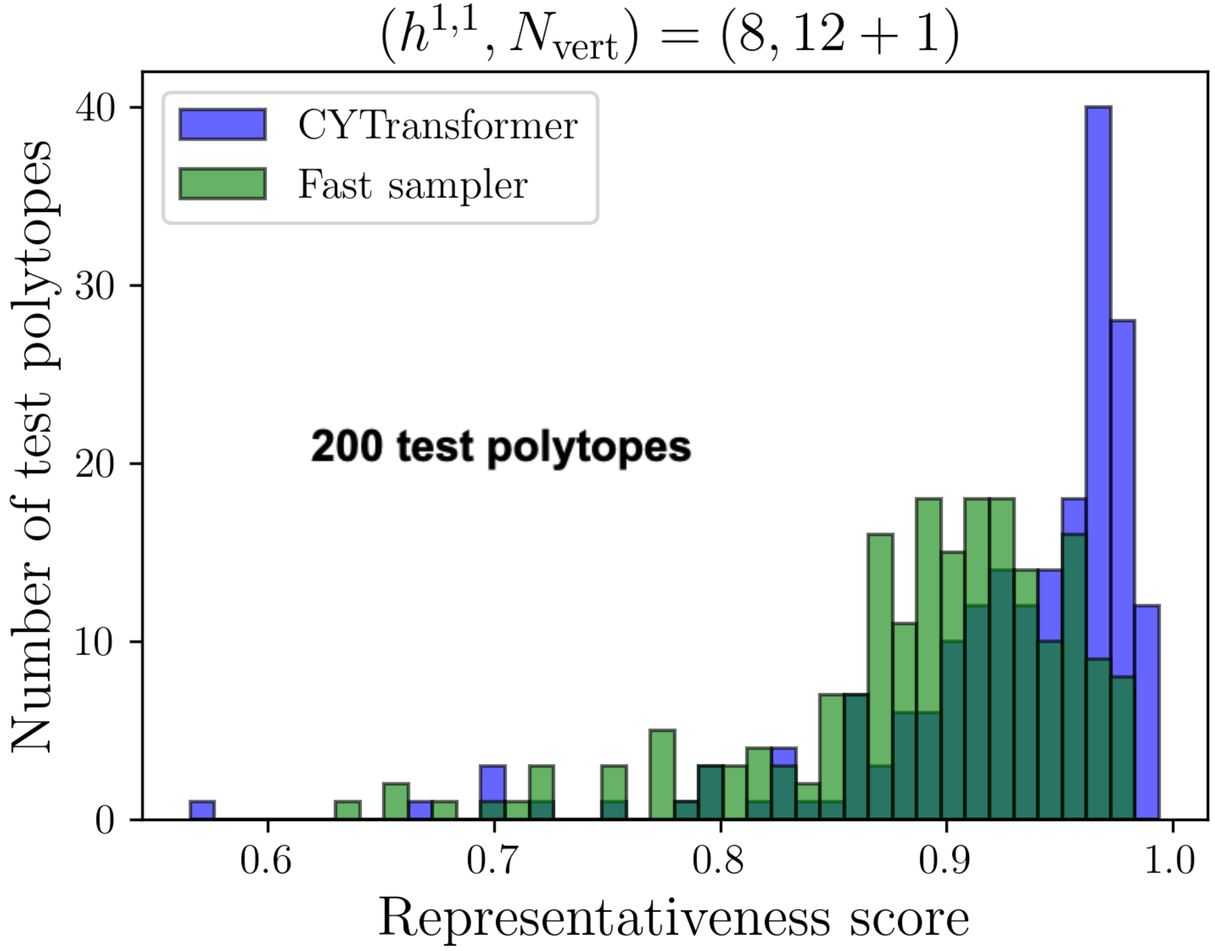
CYTransformer - Distinct FRSTs recovered

Fast sampler - Distinct FRSTs recovered

All distinct FRSTs

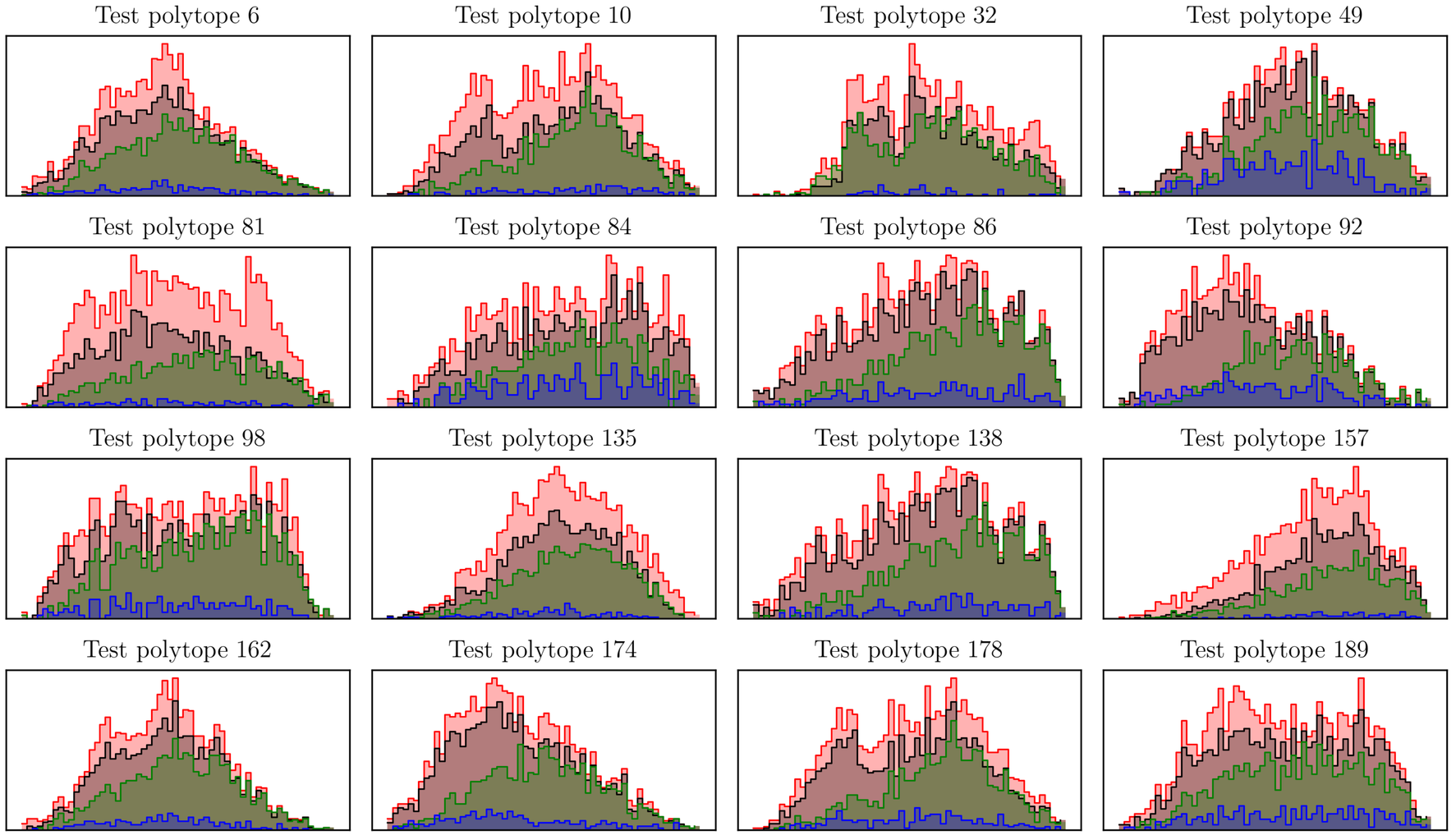
- Fast sampler does not match the population distribution - **biased sampling**
- We can **quantify the representativeness** by computing the cosine similarity between model histogram and population histogram

Comparison: Representativeness Histogram



- CYTransformer
  - Peak near 1 with low variance
  - consistent representative sampling
- Fast sampler
  - Lower, more spread out scores
  - biased sampling

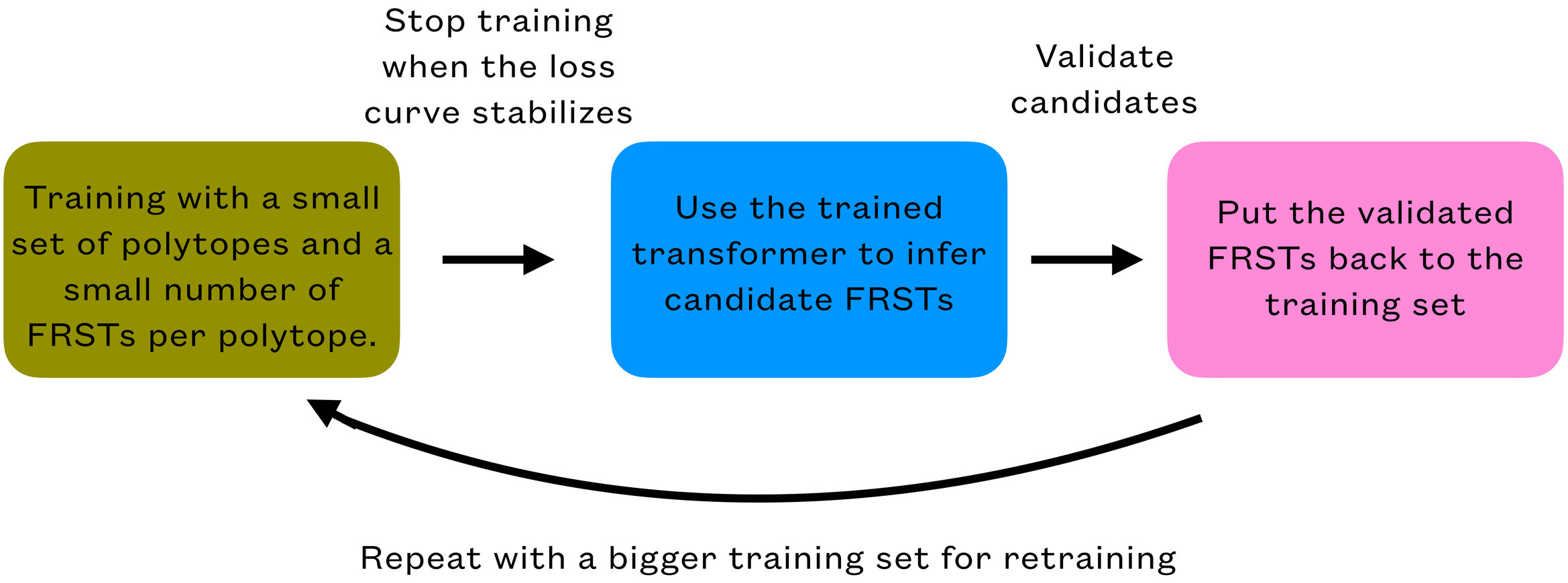
Fast and Transformative



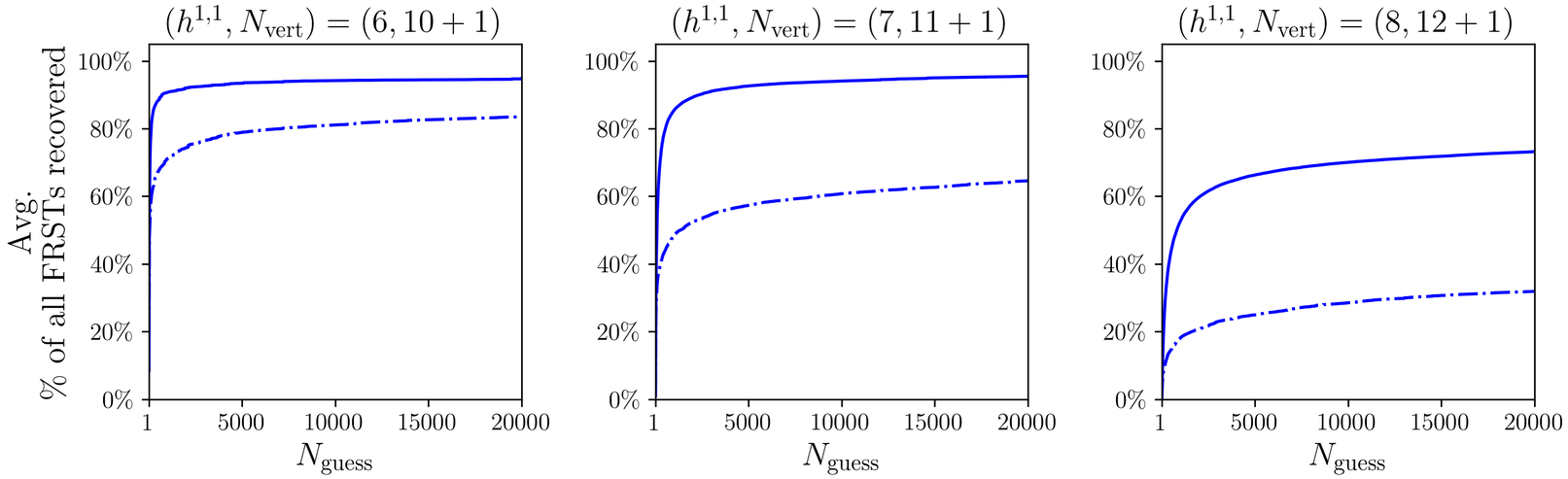
Red - Population
Blue - Seeds at N = 500
Black - Hybrid final result
Green - Pure fast sampler

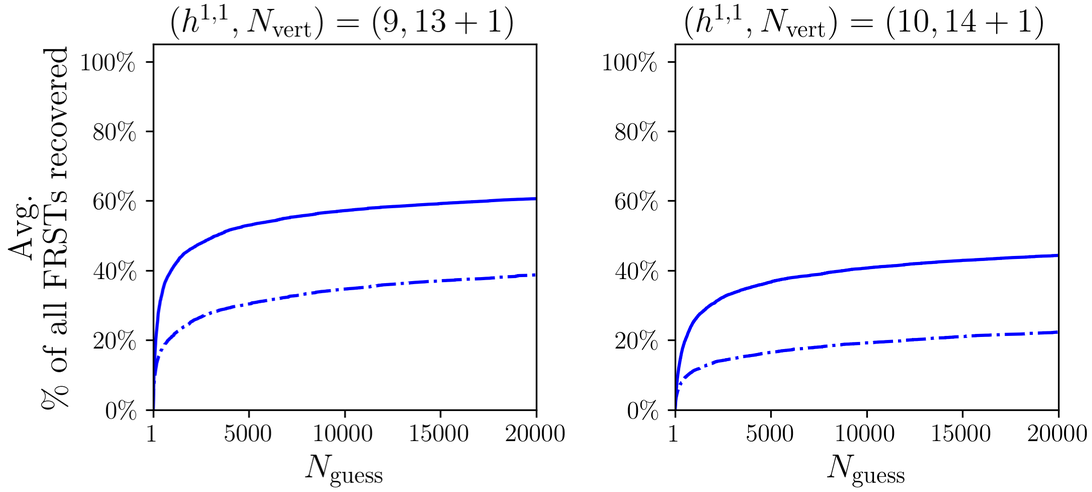
- (Not shown) Hybrid recovers more FRSTs than CYTransformer or fast sampler alone
- With just 500 CYTransformer runs for seeds, hybrid recovers the population shape
- **Hybrid maintains unbiasedness**

Self-Improvement



CYTransformer: Self-Improvement Capability





CYTransformer (after self-improvement) - Counting distinct FRSTs

CYTransformer (before self-improvement) - Counting distinct FRSTs

Exploring Heterotic Standard Models with Transformer

[Yip, Mininno, GS, to appear]

- We consider $E_8 \times E_8$ heterotic string theory general embeddings on Calabi–Yau X_3 with
  - V of E_8, holomorphic vector bundle over X_3 with structure group H, i.e., $E_8 \supset G \times H$.
  - Existence of Wilson lines to break G to the MSSM gauge group.

$$V = \bigoplus_{a=1}^{5} L_a = \bigoplus_{a=1}^{5} \mathcal{O}_{X_3}\left(\sum_{i=1}^{h^{1,1}} k_a^i J_i\right) = \mathcal{O}_{X_3}(\mathbf{k_a})$$

- Conditions:
  - Anomaly cancellation.
  - Minimal supersymmetry, i.e., poly-stablilty of the line bundles.
  - Equivariant structure: $\exists$ a freely-acting symmetry Γ on X_3 and V is a vector bundle of $\tilde{X}_3 = X_3/\Gamma$.

Transformer-based RL Explorer

[Yip, Mininno, GS, to appear]

- Assumptions we made in our search of Heterotic Standard Model :
  - Smooth and simplicial Mori cone.
  - Known action of Γ on X_3.

	Conditions	RL-Explorer
E_8 embedding	$c_1(V) = 0\,,\ \sum_{a\in S} c_1(L_a) \neq 0$	✓
Anomaly cancellation	$c_2(T\mathbf{X}_3) - c_2(V) \in$ Mori cone	✓
Poly-stability	$\mu(L_a) = 0$	✓
Three chiral families	$\mathrm{ind}(V) = -3\lvert\Gamma\rvert$	✓
No exotic representations	$-3\lvert\Gamma\rvert \leq \mathrm{ind}(L_a) \leq 0\,,\ -3\lvert\Gamma\rvert \leq \mathrm{ind}(L_a \otimes L_b) \leq 0$	✓
Equivariance*	$\sum_{\text{distinct } L\subset V} m(L)\chi(\mathbf{X}_3, L) = 0 \mod \lvert\Gamma\rvert$	✗ (Filter-out later)
Spectrum	$h^1(\mathbf{X}_3, V) = 3\lvert\Gamma\rvert\,,\quad h^2(\mathbf{X}_3, V) = 0\,,$ $h^1(\mathbf{X}_3, \wedge^2 V) = 3\lvert\Gamma\rvert + n_h\,,\quad h^2(\mathbf{X}_3, \wedge^2 V) = n_h$	✗ (Filter-out later)

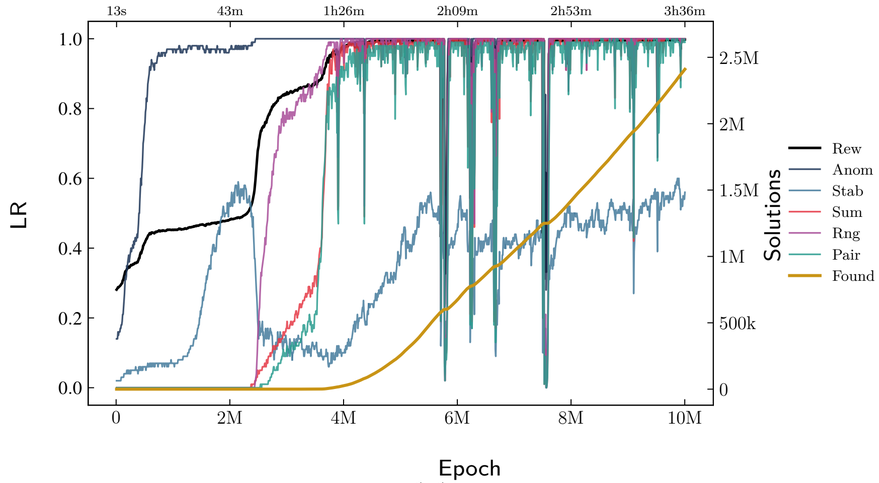
Vast literature on CICY Standard Models: [Anderson, He, Lukas, '07, '08]; [Braun, '10]; [Anderson, Gray, Lukas, Palti, '11, '12]; [Anderson, Constantin, Gray, Lukas, Palti, '13; [Lukas, Mishra, '17]; [Larfors, Schneider, '19]...

Some Results

[Yip, Mininno, GS, to appear]

- We built a flexible transformer-based RL-explorer for CY admitting smooth simplicial Mori cone and freely-acting symmetries.
- We tested it on ~40 favorable CICYs with $h^{1,1}(X_3) \in [4,15]$ and $|\Gamma| \in [2,4]$.

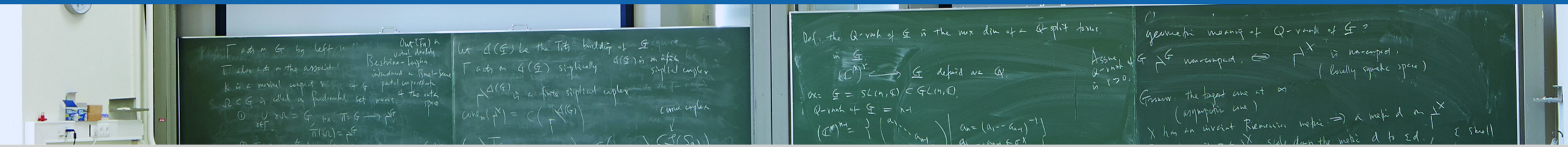
#	$h^{1,1}$	$\|\Gamma\|$	Sol.	Total	Equivariance
7447	5	2	N	43903	12670
			N_{S_5}	12658	3919
			$N_{G\times_3}$	2590	804
			N_{full}	554	204
7300	8	2	N	2403949	372913
			N_{S_5}	2365092	369556
			$N_{G\times_3}$	2289114	357146
			N_{full}	2232797	348247
5257	10	2	N	2903468	482530
			N_{S_5}	2899680	482427
			$N_{G\times_3}$	2898620	482420
			N_{full}	2878537	481878



- Exhaustic scan only up to $h^{1,1} \leq 6$ [Anderson, et al, '13]
- The versatility and scalability of our framework makes it possible to perform the deepest search (so far) and for CY with much larger $h^{1,1}$.

Summary

- Many vexing questions in **Strings and Geometry** are well-suited for AI/Machine Learning.
- We have developed CYTranformer that can be used to generate new CY 3-folds. Our method shows promise of scalability and can be continuously improved via self-improvement/priming.
- We developed a transformer-based RL explorer for Heterotic Standard Models.
- With an efficient generator of new string vacua, we can address their physics implications:
  - optimization on the Landscape: search for realistic compactifications
  - find novel predictions in particle physics and cosmology
  - discovering structure of the Landscape, and what is possible/impossible (Swampland)
  - discover new mathematical relations



The Unreasonable Effectiveness of Toric Geometry: Bridging Mathematics, Computation, and String Theory

Description Schedule Associated Events Participants

This programme aims to create a stimulating environment where mathematicians working in string inspired areas of toric geometry meet physicists who use toric geometry as a tool. Recent developments will be discussed over the disciplinary border, thereby facilitating progress on open problems and the formulation of new questions.

Topics

- Toric geometry as an arena for explorations in string theory, where a wide range of questions in physics can be tested with precise mathematical tools.
- String theory as an abundant source of intriguing mathematical questions.
- Specialist software for research in theoretical physics and mathematics (e.g. Macaulay2 and cytools) and machine learning tools for addressing open questions in theoretical physics and mathematics (e.g. ML libraries for Calabi-Yau metrics).

At a glance

Type:
Thematic Programme

When:
June 15, 2026 — July 17, 2026

Where:
ESI Boltzmann Lecture Hall

Organizer(s):
Magdalena Larfors (Uppsala U)
Gary Shiu (U of Wisconsin-Madison)
Harald Skarke (TU Vienna)
Michael E. Stillman (Cornell U, Ithaca)