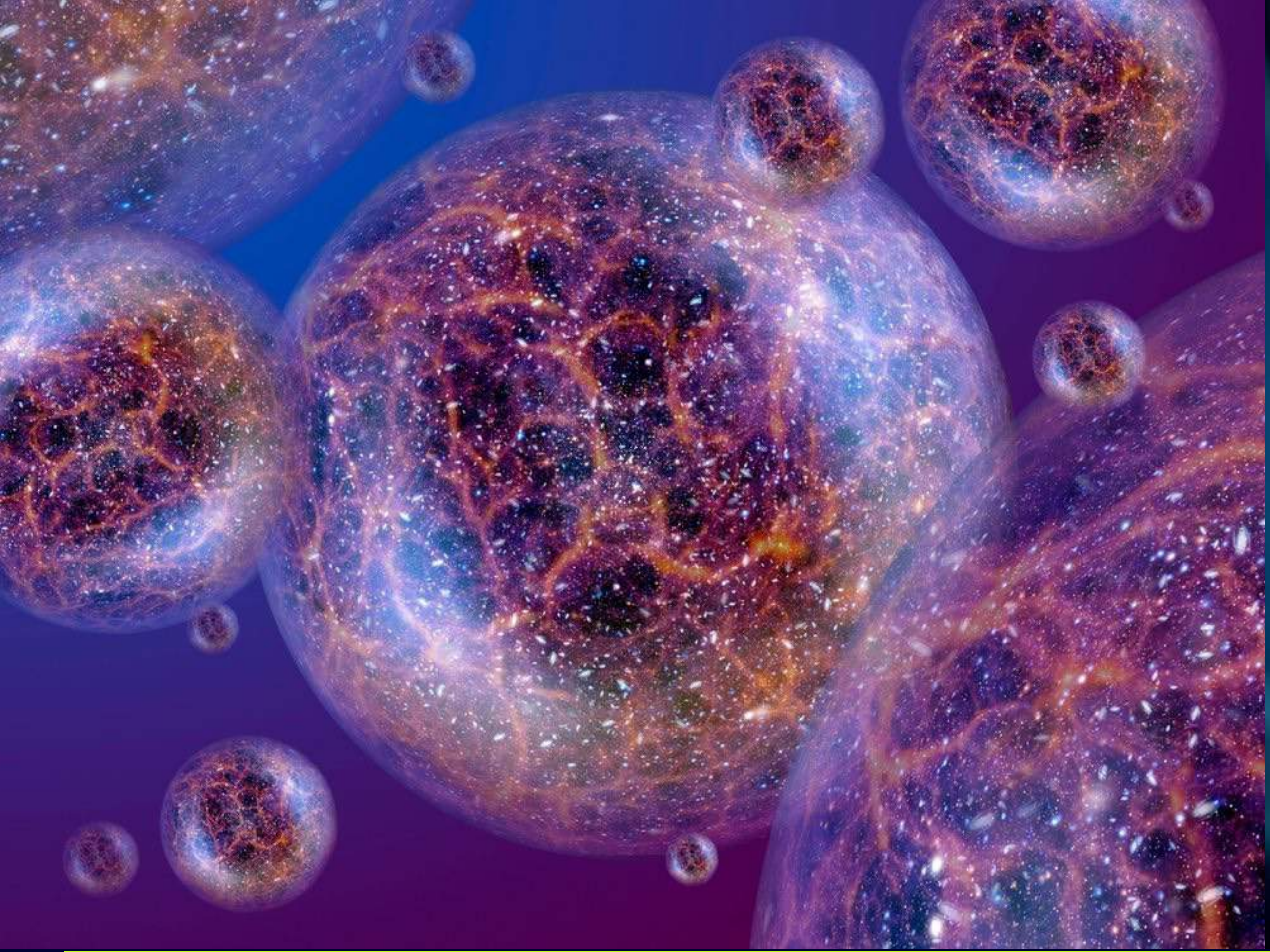


The String Genome Project

GARY SHIU, UNIVERSITY OF WISCONSIN-MADISON

The Cast

Alex Cole (Amsterdam)

Andreas Schachner (Cambridge)

Sven Krippendorf (Munich)

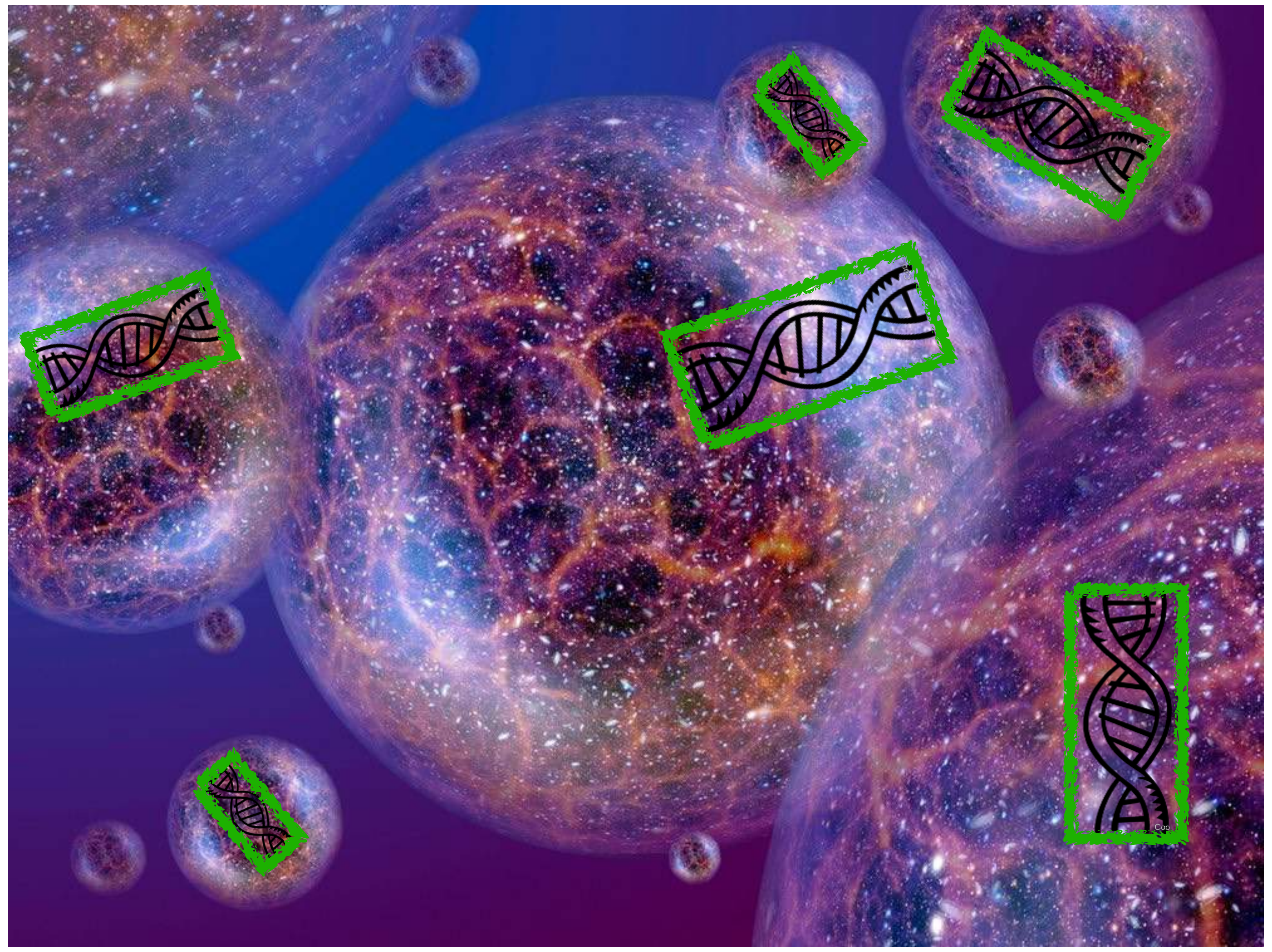
Gregory Loges (UW-Madison)

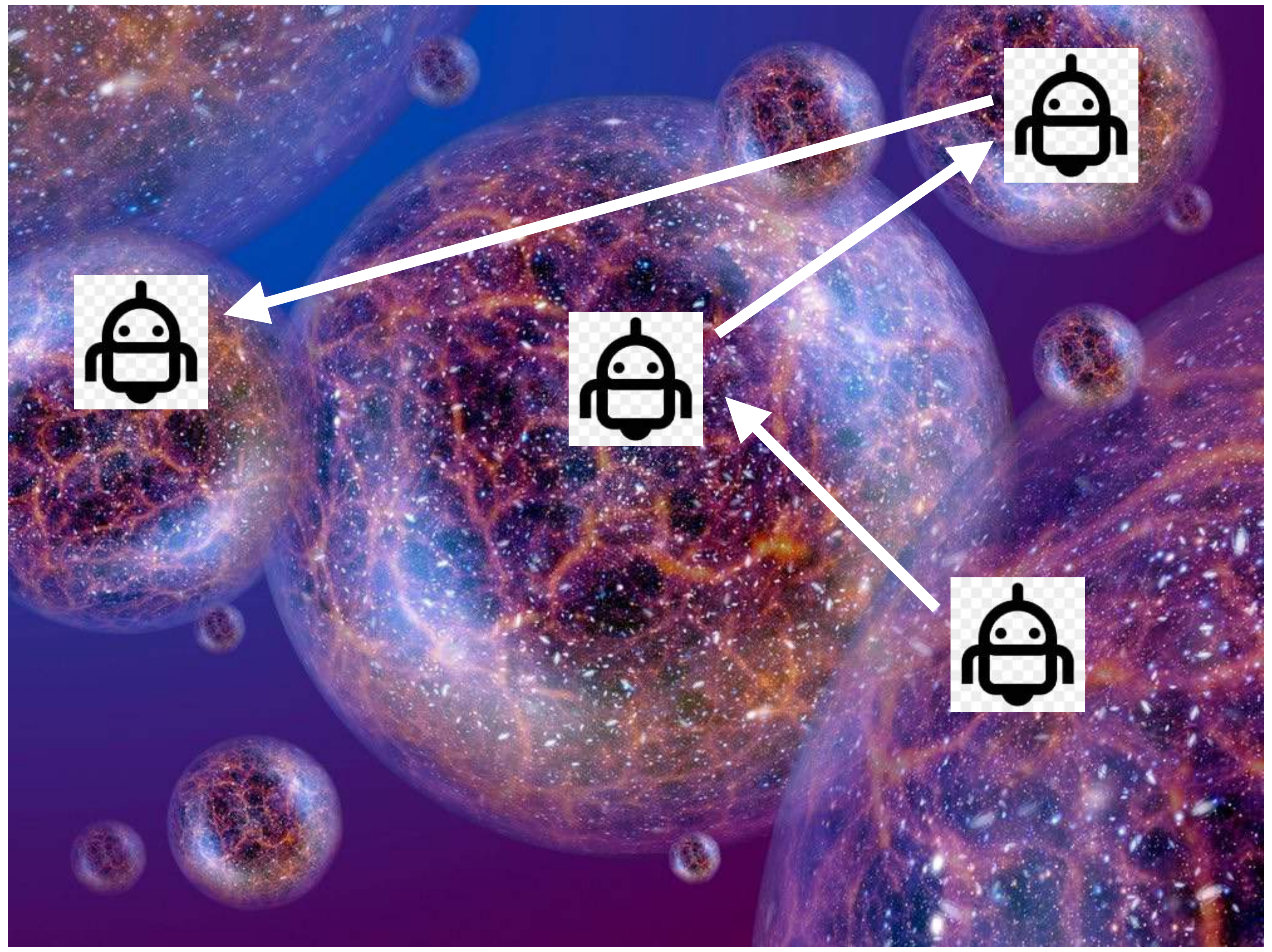
- A. Cole, A. Schachner and GS, *"Searching the Landscape of Flux Vacua with Genetic Algorithms,"* JHEP **11**, 045 (2019) [arXiv:1907.10072 [hep-th]].
- A. Cole, S. Krippendorf, A. Schachner and GS, *"Probing the Structure of String Theory Vacua with Genetic Algorithms and Reinforcement Learning,"* [arXiv:2111.11466 [hep-th]], accepted for NeurIPS 2021 Machine Learning and the Physical Sciences.
- G.J. Loges and GS, *"Breeding Realistic D-brane Models,"* [arXiv:2112.08391 [hep-th]].

Nature or Nurture?

- Do we achieve excellence by breeding or training?
- Our work does **NOT** lend support to eugenics.
- Genetics and training are **complementary**; both lead to Rome, with different traits uncovered.

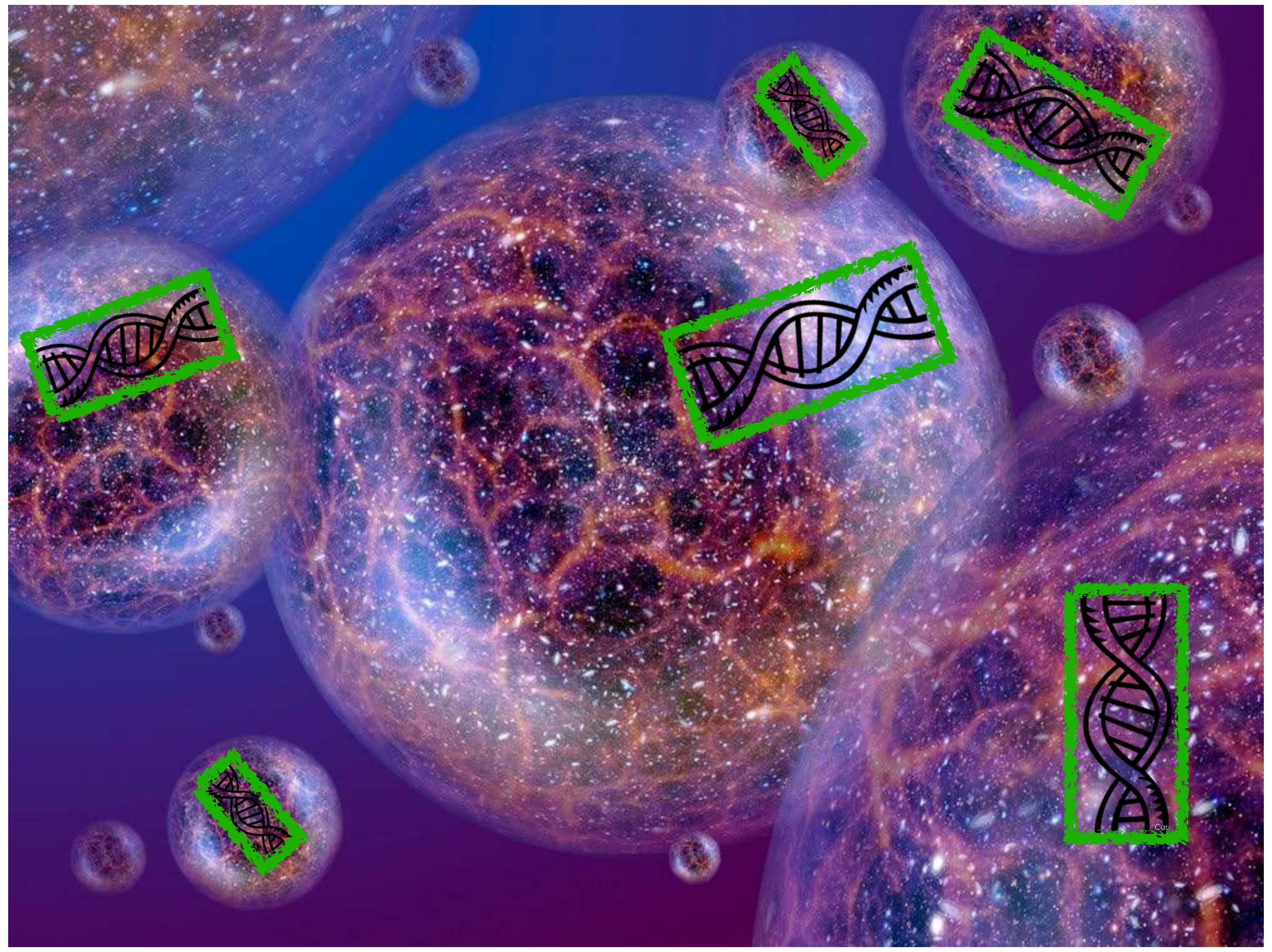
Nature or Nurture?

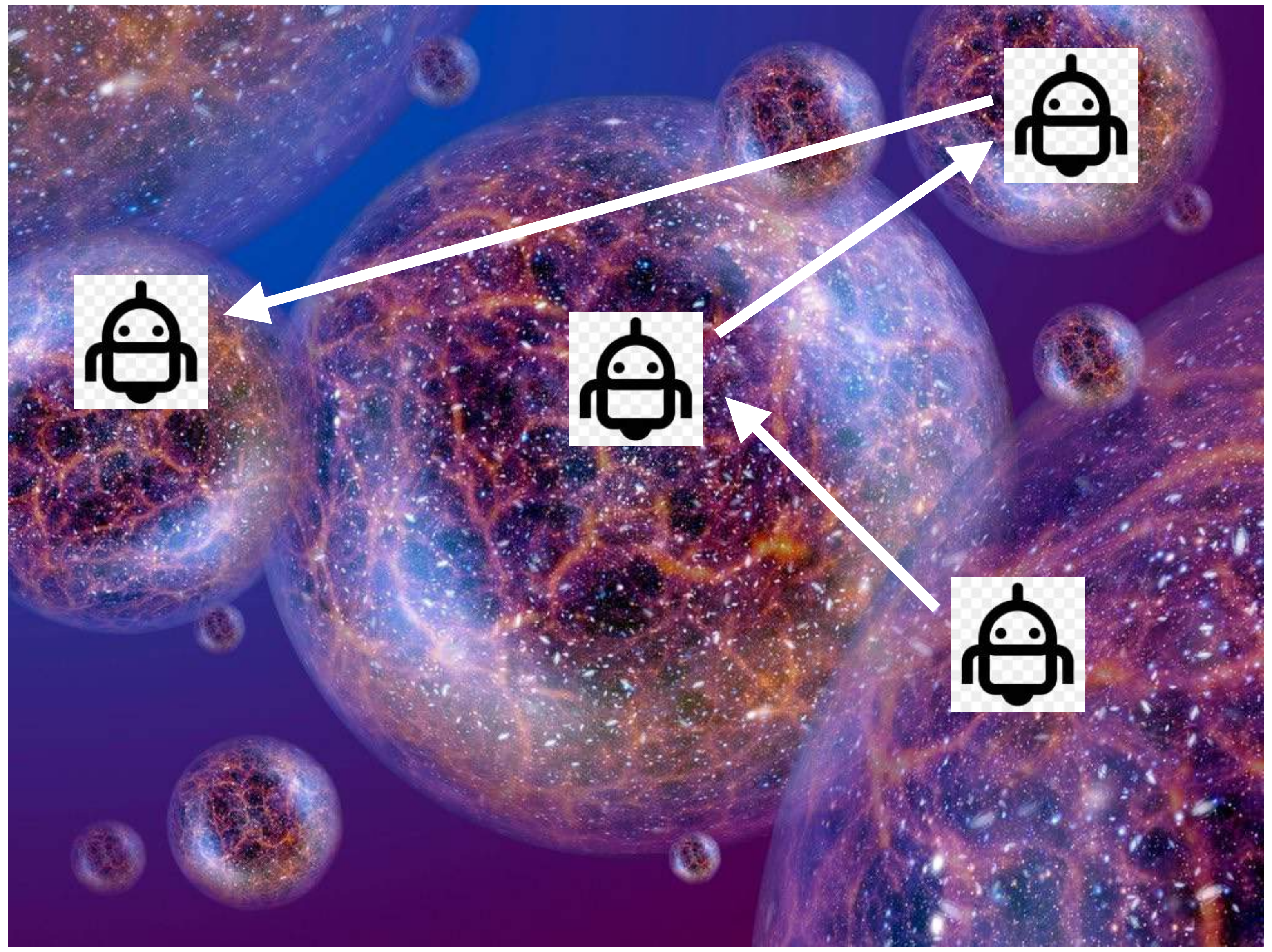




- The landscape is vast, with as many as 10^{500} [Ashok, Douglas]; [Denef, Douglas] to $10^{272,000}$ [Taylor, Wang] states, making exhaustive searches or random sampling impractical.
    - Can stochastic optimization (e.g. GAs) and RL effectively search for desirable string vacua?
    - Can these methods discover structures in the landscape? Are the structures discovered the same?

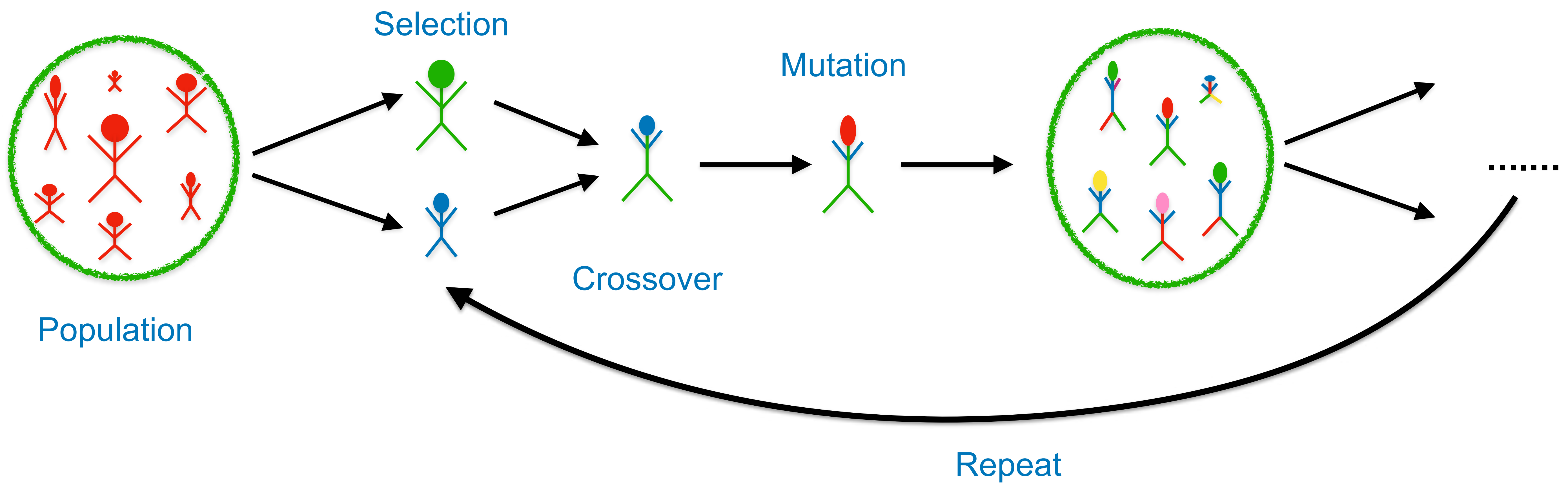
Nature or Nurture?





- Can stochastic optimization (e.g. GAs) and RL effectively search for desirable string vacua?
  - Flux vacua with desired W_0 and g_s: [Cole, Schachner, GS, '19]; [Cole, Krippendorf, Schachner, GS, '21]
  - Intersecting brane models [Loges, GS, '21]
- The above works uncover structures in the landscape. The structures found by GA and RL are complementary.

Genetic Algorithms



- Start from a (random) initial population:
    - **Selection** for breeding favors fitter individuals
    - **Breeding** via crossover (large changes).
    - **Mutations** keep the population sufficiently diverse (small changes).
    - **Elitism** ensures monotonic progress (optional).
- Rinse and Repeat.

- Many string landscape problems can be phrased as ***inverse problems*** with ***discrete variables*** (fluxes, # branes), traditional gradient-descent-like optimizations are not directly applicable.

An Example: Knapsack Problem

- The problem of finding vacua with energies within a specific range in the (toy) landscape (e.g. [Bousso, Polchinski];[Arkani-Hamed, Dimopoulos, Kachru]) is in the same universality class.

- Given a collection of items with weights and values, choose the subset which maximizes total value without exceeding the capacity of your knapsack.

Weight	5	7	8	10	14
Value	10	17	23	28	40

Capacity: 20

- An individual ζ consists of a chromosome χ describing a choice of items and a fitness which characterizes $\mathscr{F}$ how well the problem is solved:

$$\begin{array}{lll} \chi(\zeta) = 11010 & (w,v) = (22,55) & \mathcal{F}(\zeta) = 0 \\ \chi(\zeta) = 10100 & (w,v) = (13,33) & \mathcal{F}(\zeta) = 33 \\ \chi(\zeta) = 00110 & (w,v) = (18,51) & \mathcal{F}(\zeta) = 51 \end{array}$$

An Example: Knapsack Problem

- **Selection:** binary tournament to select two “parents”
- **Crossover:** combine chromosomes of two parents to create child

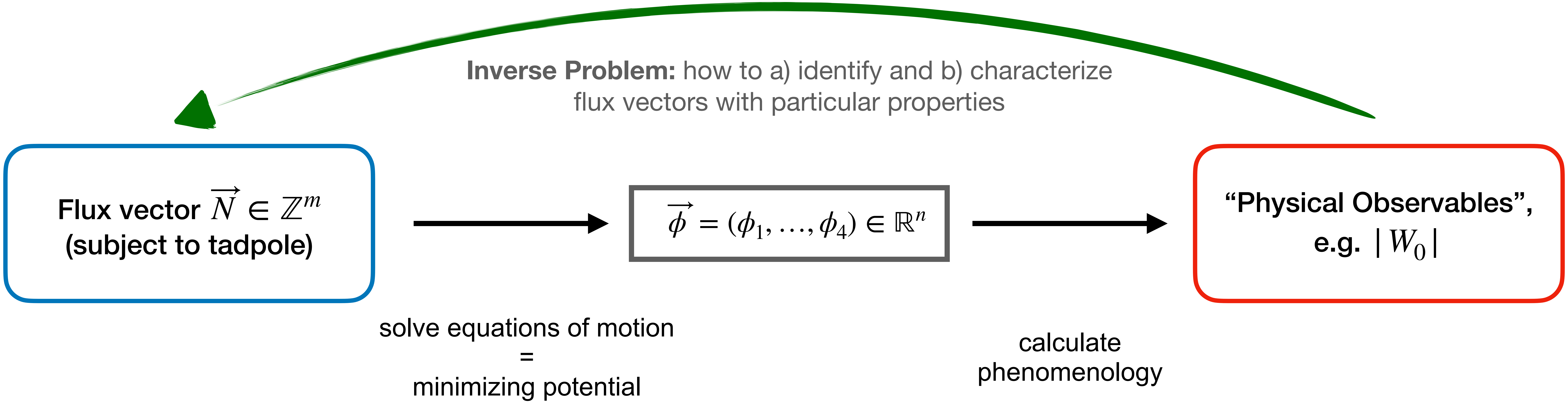
$$\begin{array}{rl} & 100101{\color{gray}110001} \\ \otimes & {\color{gray}011111}011101 \\ \hline & 100101011101 \end{array} \qquad \begin{array}{rl} & 000011100{\color{gray}101} \\ \otimes & {\color{gray}001011000}010 \\ \hline & 000011100010 \end{array}$$

- **Mutation:** flip each bit with probability p

$$110101111101 \longrightarrow 1101011{\color{red}0}1101$$

- Efficiency can be seen from a slightly scaled-up version: If we have 16 objects with certain weight and values, we can reach an optimal solution after $\mathcal{O}(10)$ generations for a population of size 50 and $\mathcal{O}(400)$ unique states visited, compared with brute-force search over $2^{16} = 65{,}536$ states.

The String Landscape



Inverse Problem: how to a) identify and b) characterize flux vectors with particular properties

Flux vector $\vec{N} \in \mathbb{Z}^m$ (subject to tadpole)

$\vec{\phi} = (\phi_1, \ldots, \phi_4) \in \mathbb{R}^n$

"Physical Observables", e.g. $|W_0|$

solve equations of motion = minimizing potential

calculate phenomenology

Challenge for string theory:

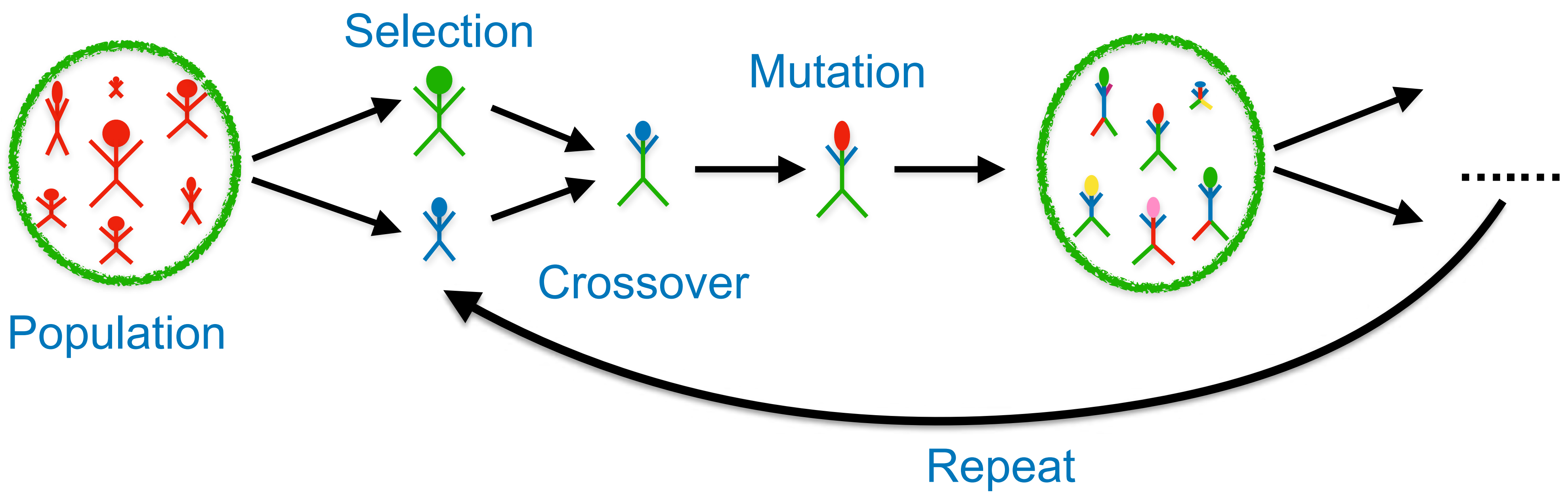
As many as $10^{272,000}$ vacua [Taylor, Wang '15], but only a few are phenomenologically interesting!

The optimization algorithms

We search for desirable string vacua via two optimization methods:

Genetic Algorithms (GAs)

Stochastic search method based on natural selection processes:

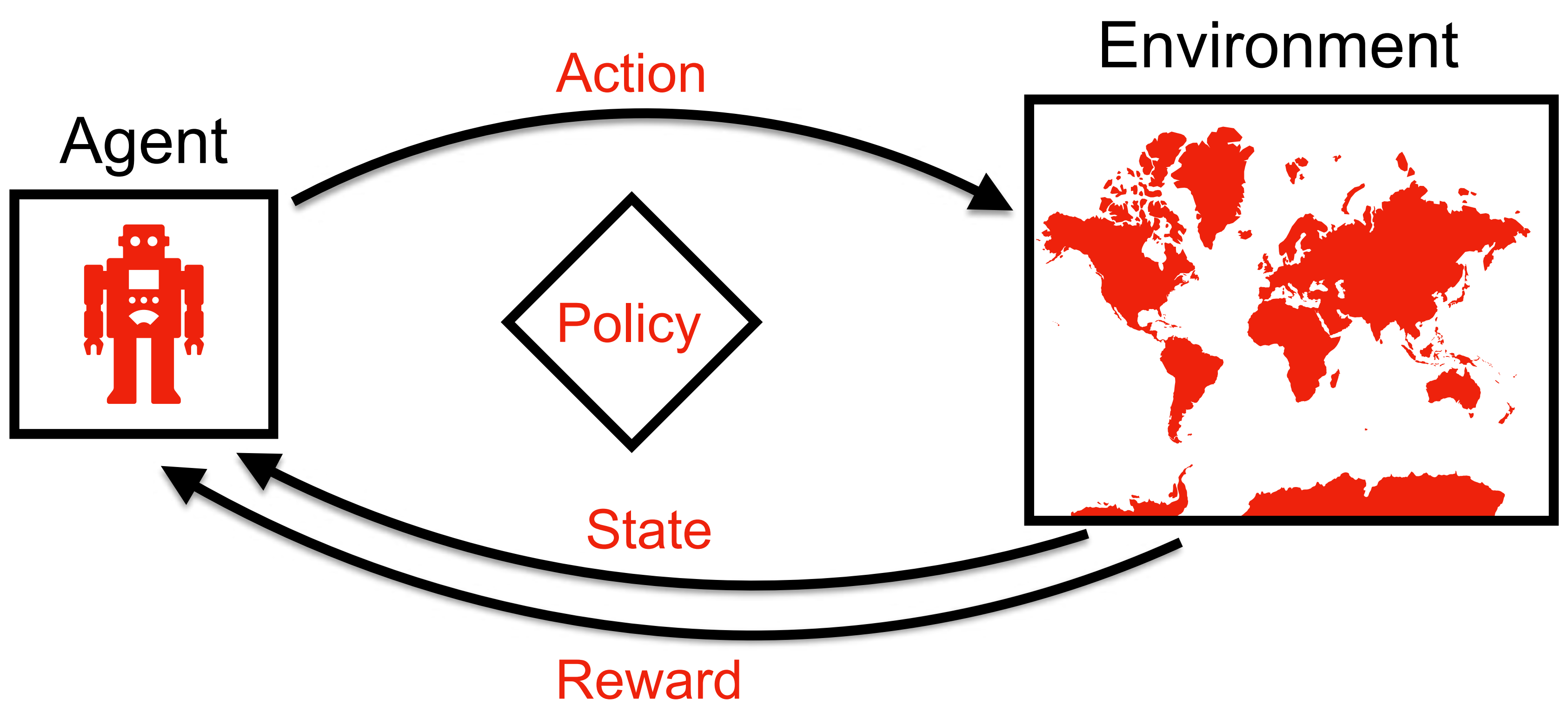


We use a generalized version of the GA of
[Cole, Schachner, GS: 1907.10072]

vs.

Reinforcement Learning (RL)

Basic ML tool to explore large environments:



We apply the A3C RL implementation of
[Krippendorf, Kroepsch, Syvaeri: 2107.04039]

We provide in [Cole, Krippendorf, Schachner, GS, ’21] the first comparison between **GAs** and **RL** in a string theory context!

See also the recent work:
[Abel et al.: 2110.14029]

To improve efficiency, we use the reward structure of [Krippendorf et al.] for RL and optimize hyperparameters in the GA by computing correlations with the mean average distance to the optimal solution.

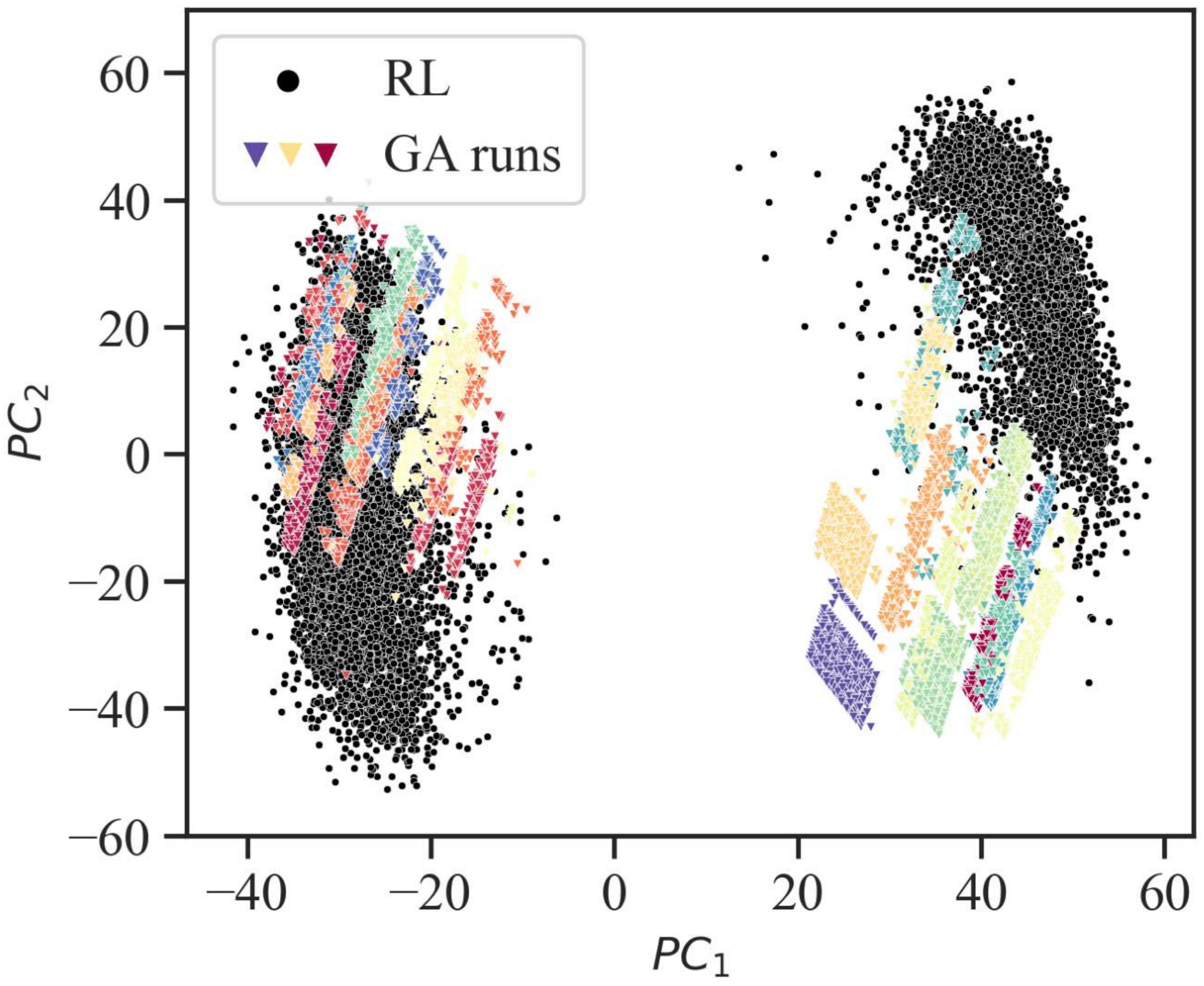
Understanding the local structure of the string landscape

Task

Apply GA+RL to find string vacua with $|W_0| = 50{,}000 \pm 1000$

We performed a **Principal Component Analysis (PCA)** on the output of flux vectors in $\mathbb{Z}^8$

PCA on combined output



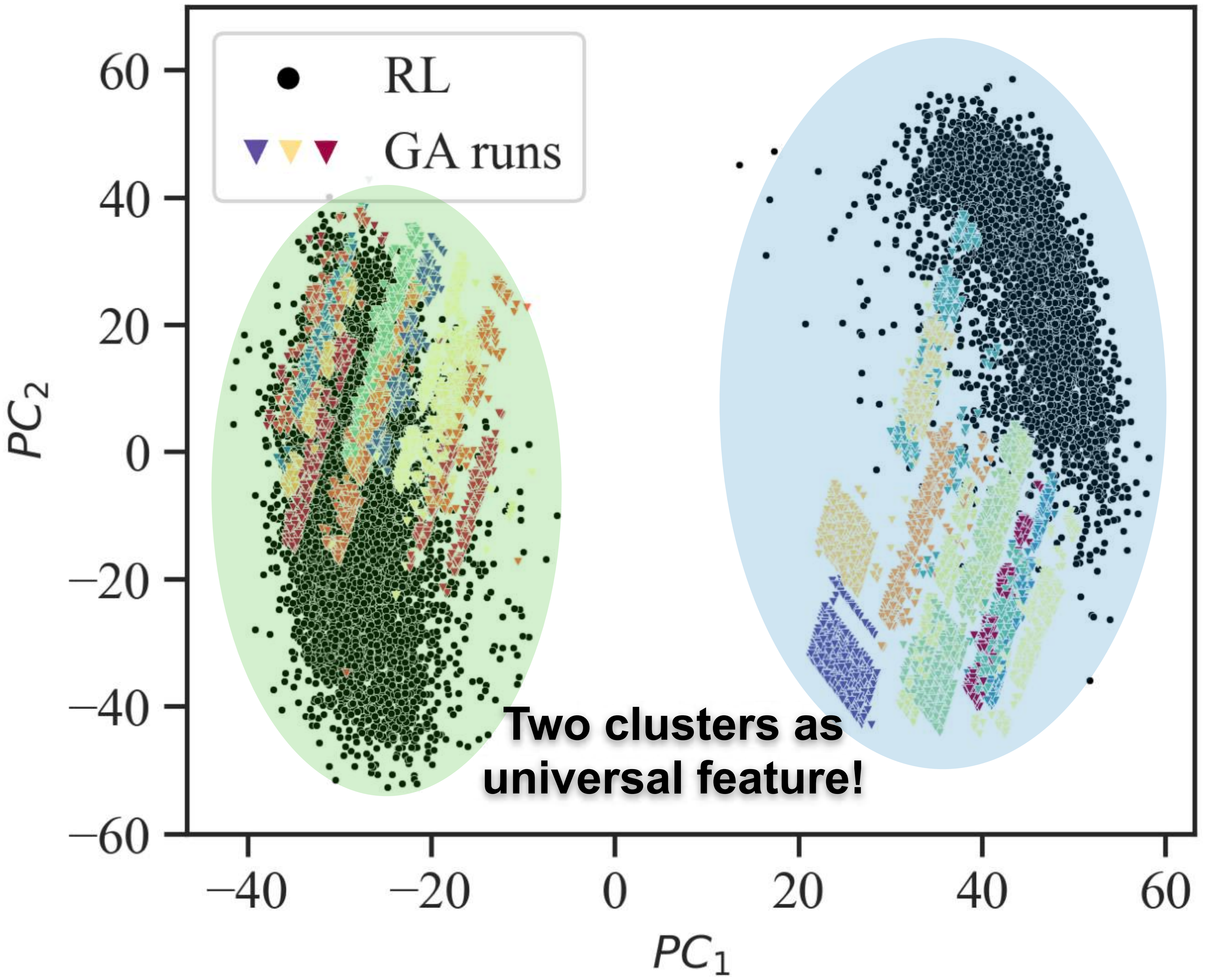
Understanding the local structure of the string landscape

Task

Apply GA+RL to find string vacua with $|W_0| = 50{,}000 \pm 1000$

We performed a **Principal Component Analysis (PCA)** on the output of flux vectors in $\mathbb{Z}^8$

PCA on combined output



Two clusters as universal feature!

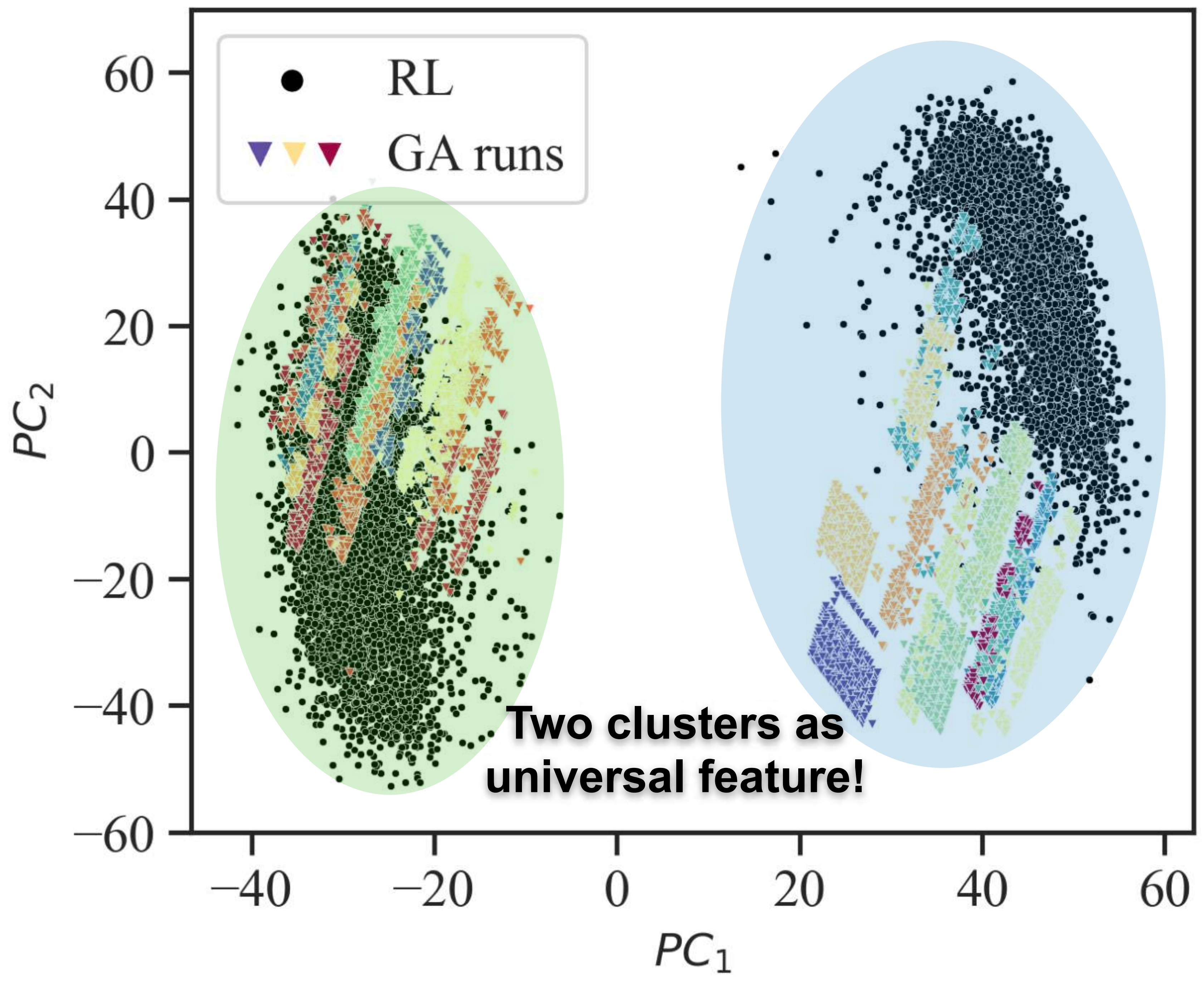
Understanding the local structure of the string landscape

Task

Apply GA+RL to find string vacua with $|W_0| = 50{,}000 \pm 1000$

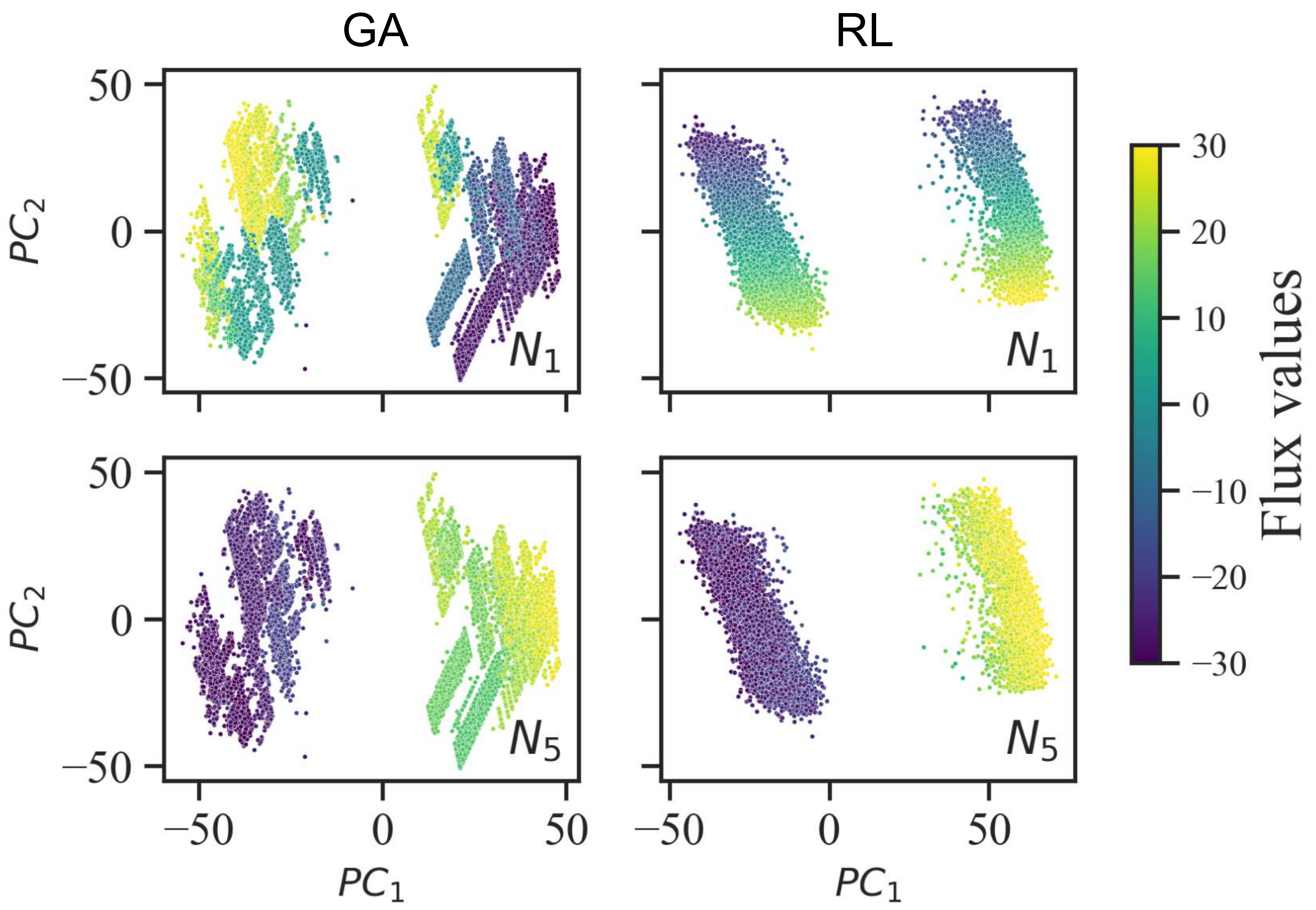
We performed a **Principal Component Analysis (PCA)** on the output of flux vectors in $\mathbb{Z}^8$

PCA on combined output



Two clusters as universal feature!

PCA on individual output



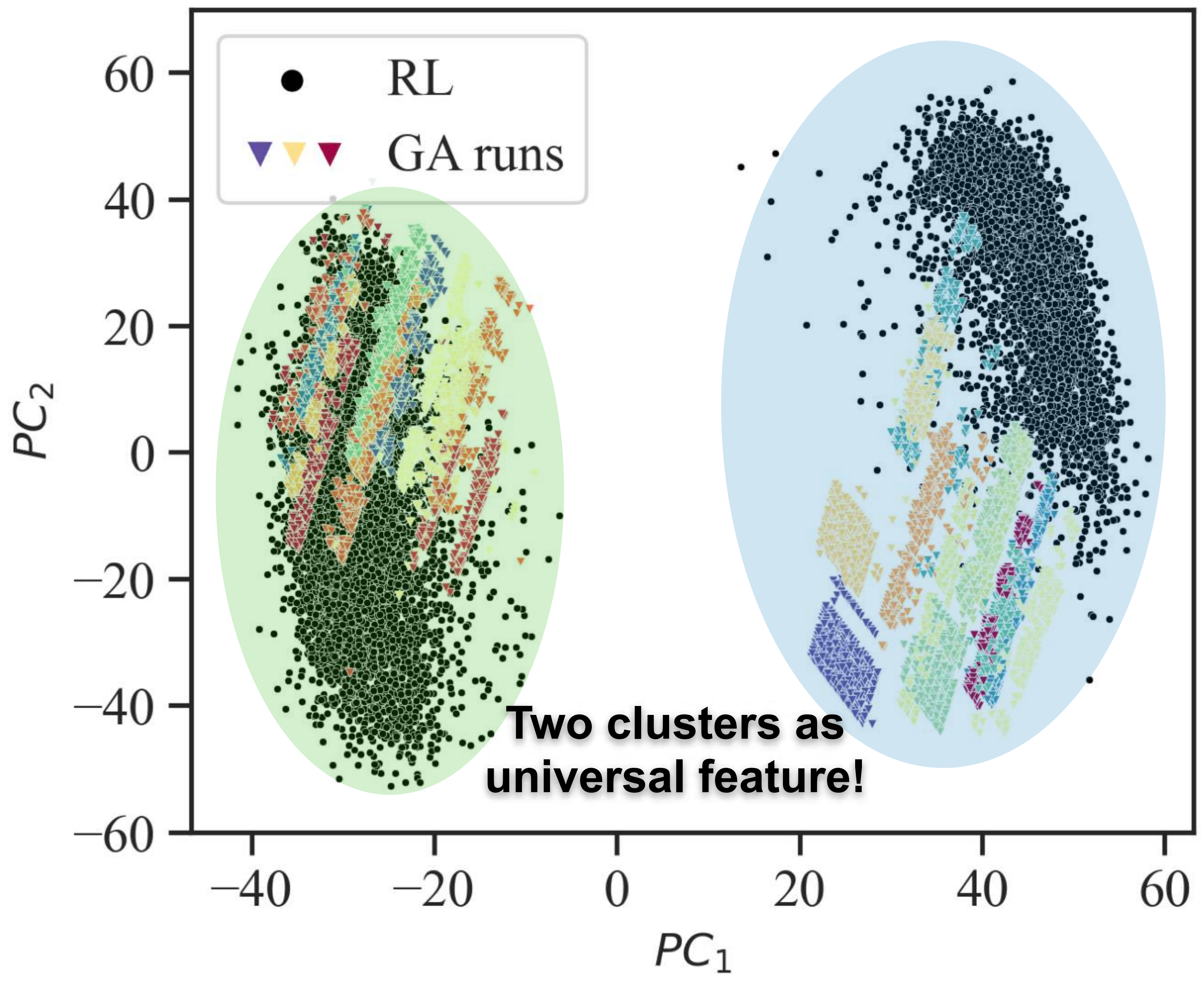
Understanding the local structure of the string landscape

Task

Apply GA+RL to find string vacua with $|W_0| = 50{,}000 \pm 1000$

We performed a **Principal Component Analysis (PCA)** on the output of flux vectors in $\mathbb{Z}^8$

PCA on combined output

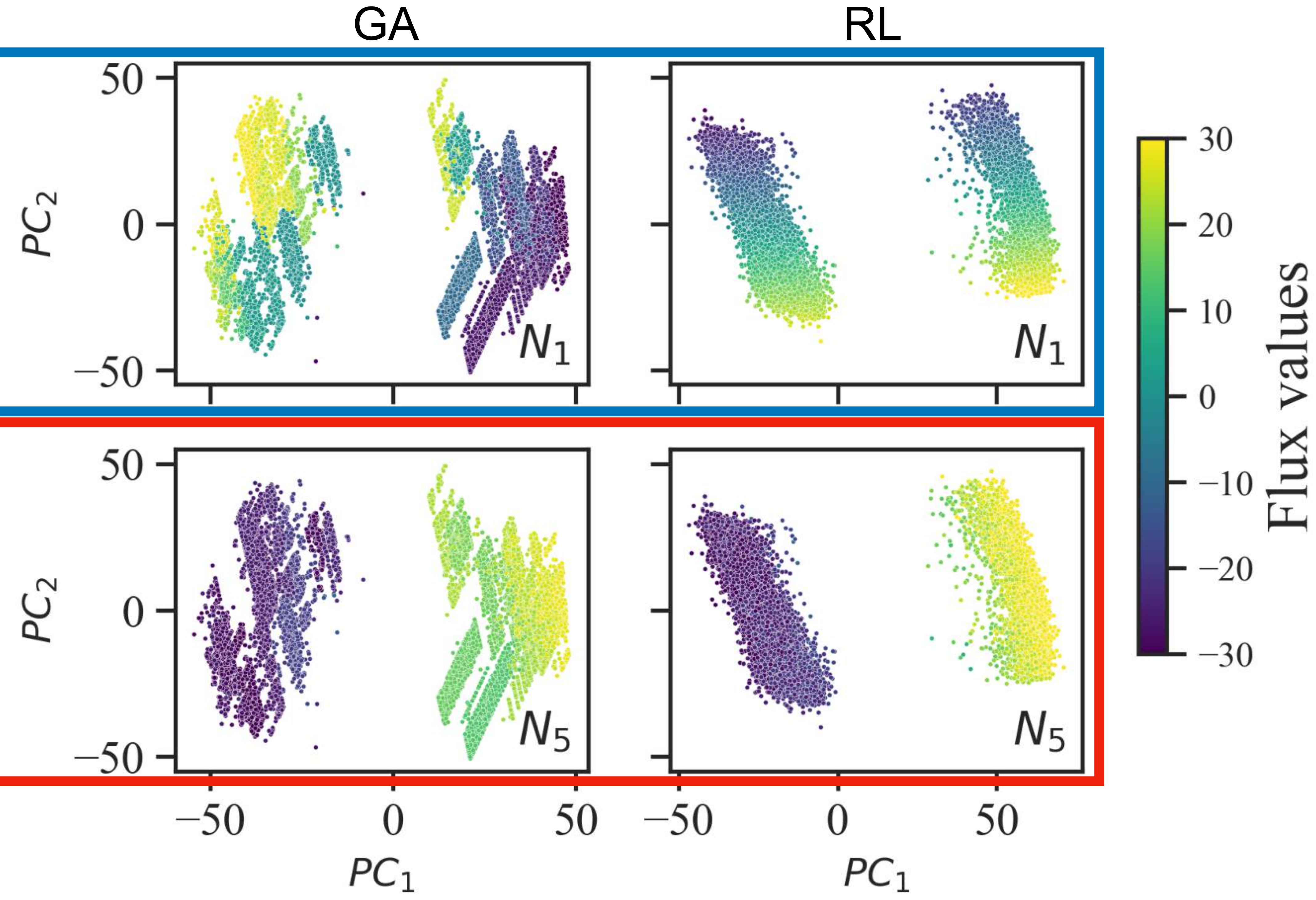


Two clusters as universal feature!

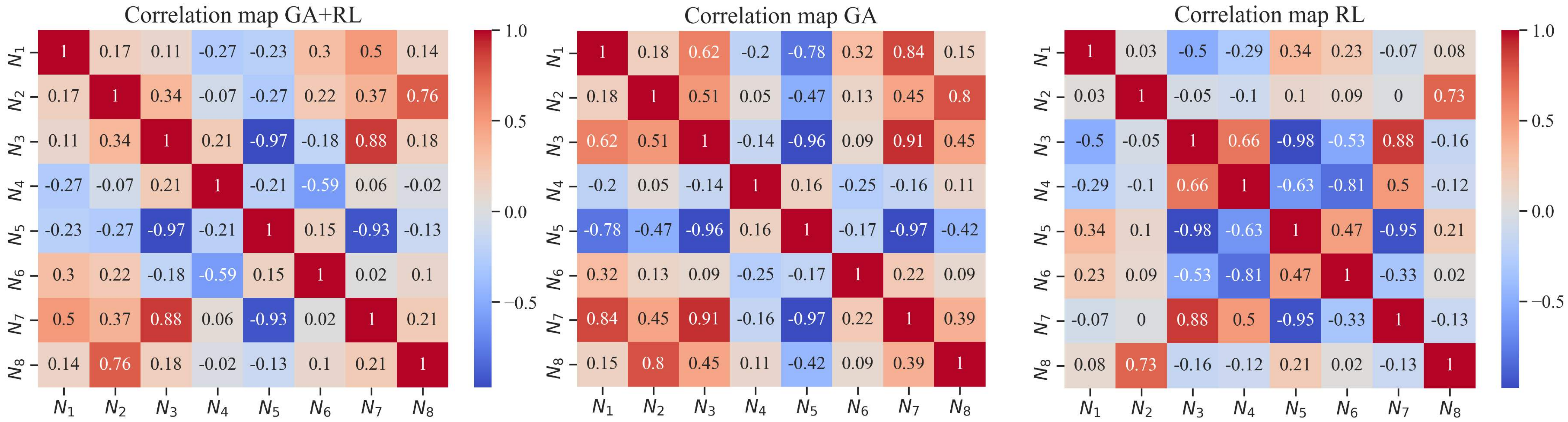
PCA on individual output

Scaling with respect to N_1 reveals difference in GA/RL output

Hint for discrete symmetry $N_5 \to -N_5$



Correlations on the Landscape



Correlation map GA+RL

	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8
N_1	1	0.17	0.11	-0.27	-0.23	0.3	0.5	0.14
N_2	0.17	1	0.34	-0.07	-0.27	0.22	0.37	0.76
N_3	0.11	0.34	1	0.21	-0.97	-0.18	0.88	0.18
N_4	-0.27	-0.07	0.21	1	-0.21	-0.59	0.06	-0.02
N_5	-0.23	-0.27	-0.97	-0.21	1	0.15	-0.93	-0.13
N_6	0.3	0.22	-0.18	-0.59	0.15	1	0.02	0.1
N_7	0.5	0.37	0.88	0.06	-0.93	0.02	1	0.21
N_8	0.14	0.76	0.18	-0.02	-0.13	0.1	0.21	1

Correlation map GA

	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8
N_1	1	0.18	0.62	-0.2	-0.78	0.32	0.84	0.15
N_2	0.18	1	0.51	0.05	-0.47	0.13	0.45	0.8
N_3	0.62	0.51	1	-0.14	-0.96	0.09	0.91	0.45
N_4	-0.2	0.05	-0.14	1	0.16	-0.25	-0.16	0.11
N_5	-0.78	-0.47	-0.96	0.16	1	-0.17	-0.97	-0.42
N_6	0.32	0.13	0.09	-0.25	-0.17	1	0.22	0.09
N_7	0.84	0.45	0.91	-0.16	-0.97	0.22	1	0.39
N_8	0.15	0.8	0.45	0.11	-0.42	0.09	0.39	1

Correlation map RL

	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8
N_1	1	0.03	-0.5	-0.29	0.34	0.23	-0.07	0.08
N_2	0.03	1	-0.05	-0.1	0.1	0.09	0	0.73
N_3	-0.5	-0.05	1	0.66	-0.98	-0.53	0.88	-0.16
N_4	-0.29	-0.1	0.66	1	-0.63	-0.81	0.5	-0.12
N_5	0.34	0.1	-0.98	-0.63	1	0.47	-0.95	0.21
N_6	0.23	0.09	-0.53	-0.81	0.47	1	-0.33	0.02
N_7	-0.07	0	0.88	0.5	-0.95	-0.33	1	-0.13
N_8	0.08	0.73	-0.16	-0.12	0.21	0.02	-0.13	1

- Some correlations are obvious as pairs of fluxes contribute as products to the tadpoles.
- There are correlations unrelated to the tadpoles.
- Comparing individual correlation maps with the combined one can unpack how GA & RL find solutions.

Take Away Messages

(1) We demonstrated that GAs and RL are efficient in finding desirable string vacua.

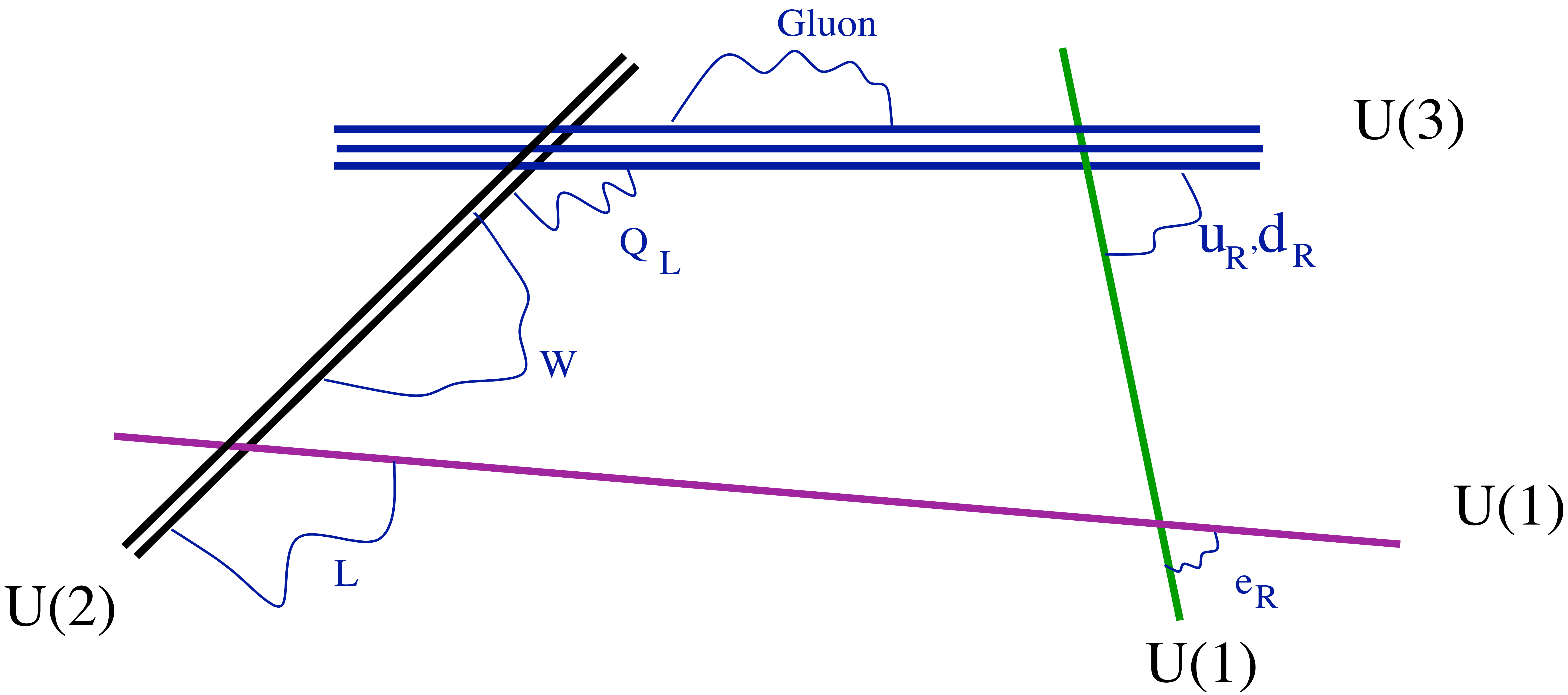
(2) Combining the output of GAs and RL allows to identify universal structures in the string landscape due to reduced sampling bias.

(3) We provided evidence for previously unknown symmetries and correlations among special string vacua.

Future directions

- Apply to tasks where reward structure becomes sparse such as $|W_0| \ll 1$.
- Determine structure behind and constraints on string theory solutions more systematically.
- Investigate more complex backgrounds and different phenomenological requirements.

Breeding Realistic D-brane Models



- ***Inverse problem*** of finding what discrete choices (#branes, how they wrap the internal space) would give phenomenologically desirable vacua (spectrum, couplings etc) [Loges, GS, '21].

Intersecting Brane Models

- Type IIA orientifolds with intersecting D6-branes (see [Blumenhagen, Cvetic, GS, Langacker, '05];[Blumenhagen, Kors, Lust, Stieberger, '07]; [Ibanez & Uranga's book, '12] for reviews) have desirable built-in features:
  - Non-Abelian gauge groups on branes
  - (Replicated) Chiral matter at brane intersections
- Timeline for the "harmonic oscillator" of intersecting brane models i.e. Type IIA on $T^6/\mathbb{Z}_2 \times \mathbb{Z}_2$:
  - $\mathcal{N} = 1$, 4D compact models [Cvetic, GS, Uranga, '01]x2
  - "One in a billion" estimate [Gmeiner, Blumenhagen, Honecker, Lust, Weigand, '06]
  - Proof of finiteness [Douglas, Taylor, '06]
  - Reinforcement learning [Halverson, Nelson, Ruehle, '19]
  - Genetic Algorithm [Loges, GS, '21]

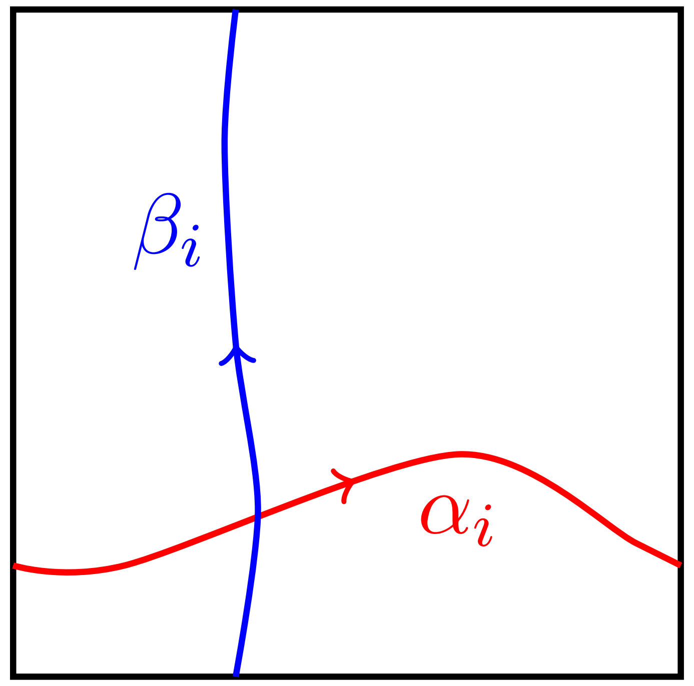
Type IIA Orientifolds

- Type IIA orientifolds of Calabi-Yau X: Weak coupling duals of G_2 compactifications of M-theory.
  [See Lukas's talk for investigation on heterotic models]

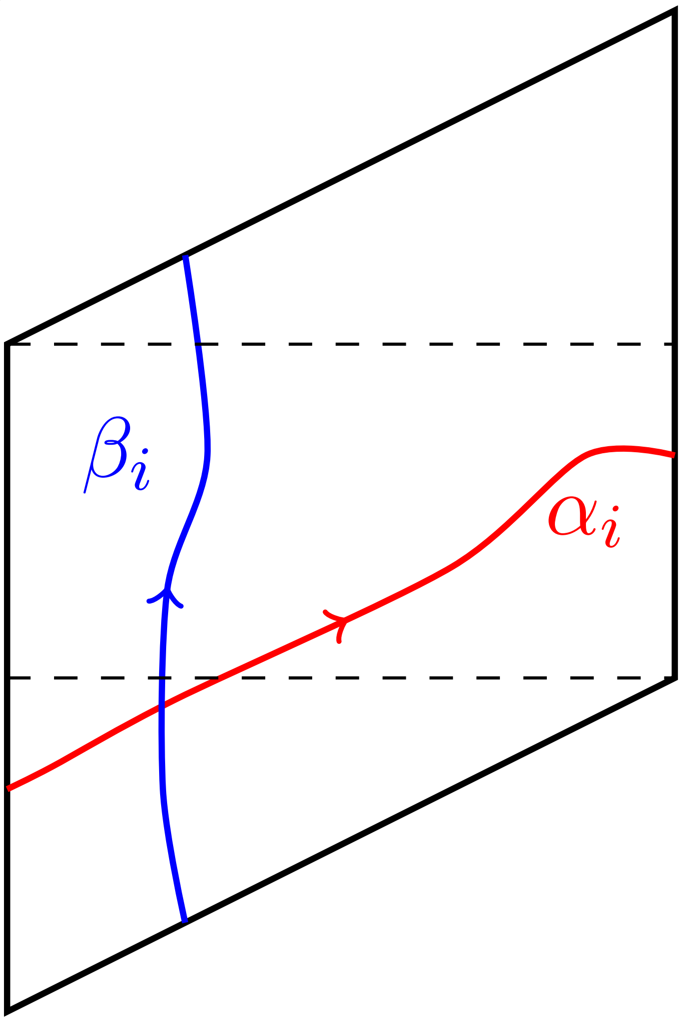
- Consider $X = T^6/\mathbb{Z}_2 \times \mathbb{Z}_2$ where
$$\begin{aligned} \theta : \quad & (z_1, z_2, z_3) \mapsto (z_1, -z_2, -z_3)\,, \\ \omega : \quad & (z_1, z_2, z_3) \mapsto (-z_1, z_2, -z_3)\,. \end{aligned}$$

- Orientifold action $\Omega\overline{\sigma}(-1)^F$ with $\overline{\sigma} : z_i \mapsto \bar{z}_i$

- Two torus shapes are compatible with $\overline{\sigma}$:

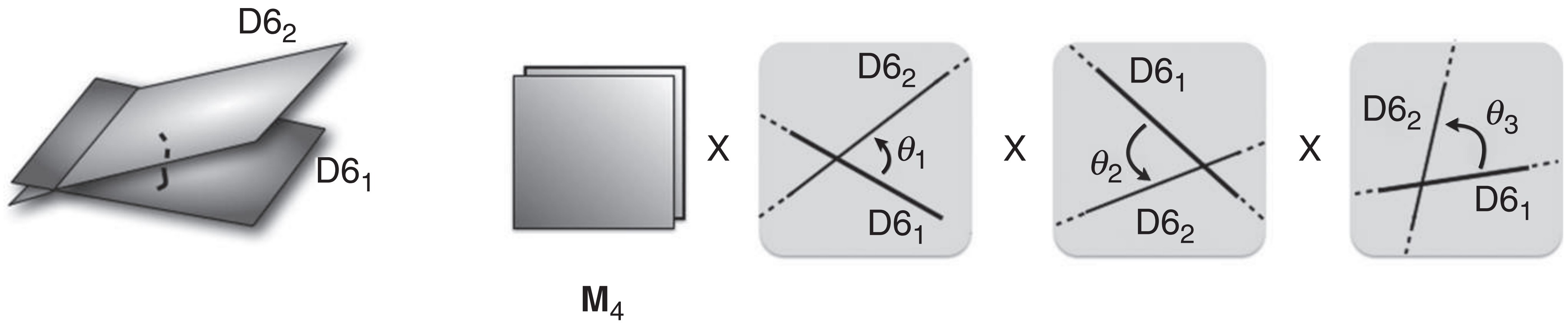


$b_i = 0$



$b_i = \frac{1}{2}$

Factorized Cycles



- 3-cycles on which the D6 wrap are specified by three pairs of **co-prime** winding numbers.

- Expand in a basis of **orientifold-even** and **-odd cycles**:

$$[\Pi_a] = \bigotimes_{i=1}^{3} \left(n_a^i[\alpha_i] + m_a^i[\beta_i]\right) = \sum_{I=0}^{3} \left(\widehat{X}{}_a^I[\pi_I^+] + \widehat{Y}{}_a^I[\pi_I^-] \right)$$

$$\begin{array}{lll} [\pi_0^+] = [\hat{\alpha}_1][\hat{\alpha}_2][\hat{\alpha}_3] & \widehat{X}{}_a^0 = n_a^1 n_a^2 n_a^3 & \widehat{Y}{}_a^0 = \widehat{m}{}_a^1 \widehat{m}{}_a^2 \widehat{m}{}_a^3 \\ [\pi_1^+] = -[\hat{\alpha}_1][\beta_2][\beta_3] & \widehat{X}{}_a^1 = -n_a^1 \widehat{m}{}_a^2 \widehat{m}{}_a^3 & \widehat{Y}{}_a^1 = -\widehat{m}{}_a^1 n_a^2 n_a^3 \\ \vdots & \vdots & \vdots \end{array}$$

with $\hat{\alpha}_i = \alpha_i - b_i\beta_i$ and $\widehat{m}{}_a^i = \dfrac{m_a^i + b_i n_a^i}{1 - b_i} \in \mathbb{Z}$.

- Individuals' **chromosomes** will consist of **stack size** and **winding numbers**.

Consistency Conditions

- **Tadpole cancellation:**

$$\sum_a N_a \widehat{X}_a^I = 8$$

- **K-theory constraints** (absence of global anomalies):

$$\sum_a N_a \widehat{Y}_a^I \in 2\mathbb{Z}$$

- **Supersymmetry** (compatibility with O6-planes):

$$\sum_{I=0}^{3} \widehat{X}_a^I \widehat{U}_I > 0 \qquad \sum_{I=0}^{3} \frac{\widehat{Y}_a^I}{\widehat{U}_I} = 0 \qquad \widehat{U}_0 = R_x^1 R_x^2 R_x^3 \quad \text{etc.}$$

- Lemma [Loges, GS, '21]: we can take $\widehat{U}_I$ to be **positive co-prime integers** WLOG to greatly simplify the search for solutions.

Brane Classification

SUSY branes classified in [Douglas, Taylor, '06]

Class	$\widehat{X}^I$	$\widehat{Y}^I$
A	$(+,+,+,-)$	$(\pm,\pm,\pm,\mp)$
B	$(+,+,0,0)$	$(0,0,\pm,\mp)$
C	$(+,0,0,0)$	$(0,0,0,0)$
A'	$(-,-,-,+)$	$(\pm,\pm,\pm,\mp)$
B'	$(+,-,0,0)$	$(0,0,\pm,\pm)$
C'	$(-,0,0,0)$	$(0,0,0,0)$
D'	$(\pm,0,0,0)$	$(0,\pm,0,0)$
E'	$(0,0,0,0)$	$(\pm,0,0,0)$

"filler branes" coined by [Cvetic, Papadimitriou, GS, '02]:

i) preserves SUSY for all complex structure moduli,

ii) satisfies K-theory constraint,

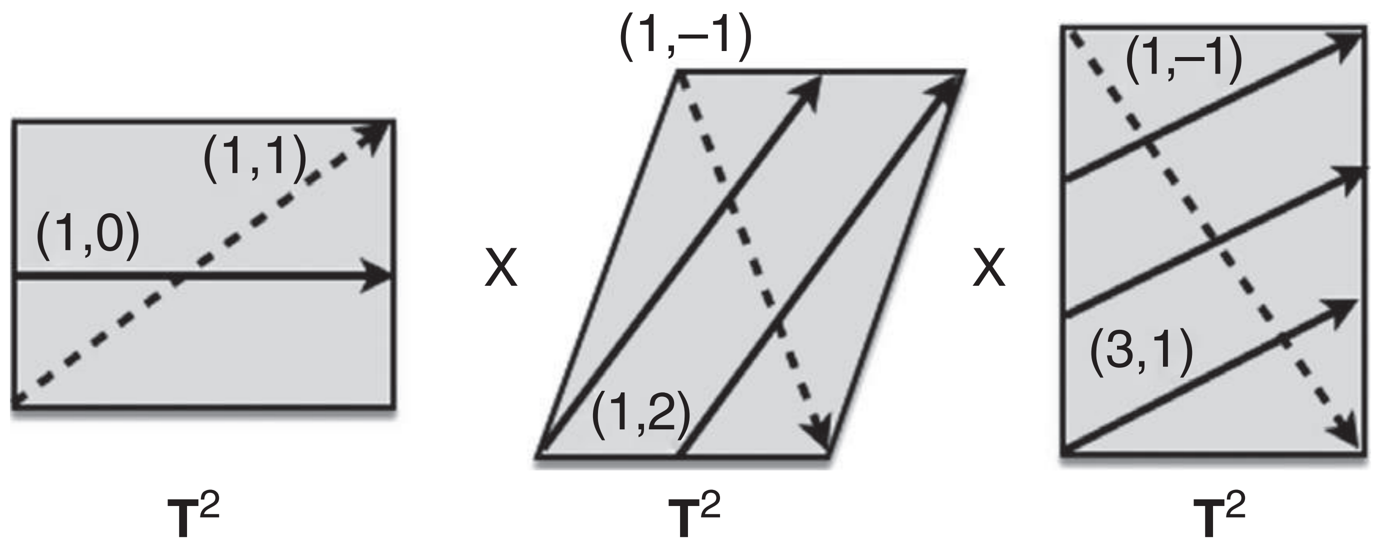
iii) contribute positively to only one tadpole.

Phenomenological Properties

- Each stack of N_a D6-branes: $U(N_a)$ gauge group in general but $USp(N_a)$ if $\widehat{Y}^I_a = 0\ \forall I$ **(C-branes)**

- Intersection of stack a and b are replicated $I_{ab} = [\Pi_a] \circ [\Pi_b]$ times:

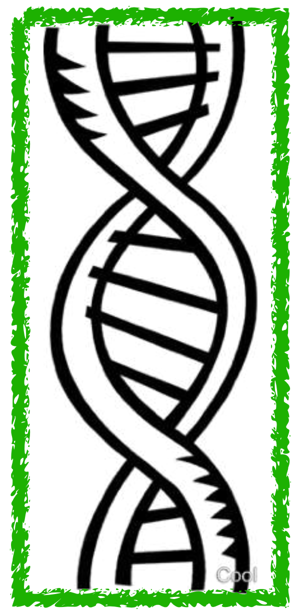
$$I_{ab} = \prod_{i=1}^{3}(n_a^i m_b^i - m_a^i n_b^i) = \prod_{i=1}^{3}(1-b_i) \times \sum_{I=0}^{3}\left(\widehat{X}^I_a \widehat{Y}^I_b - \widehat{Y}^I_a \widehat{X}^I_b\right)$$



- **Chiral spectrum**:

Representation	$(\mathbf{N}_a, \overline{\mathbf{N}}_b)$	$(\mathbf{N}_a, \mathbf{N}_b)$	$\Box\!\Box_a$	$\begin{smallmatrix}\Box\\ \Box\end{smallmatrix}_a$
Multiplicity	I_{ab}	$I_{ab'}$	$\frac{1}{2}(I_{aa'} - I_{aO6})$	$\frac{1}{2}(I_{aa'} + I_{aO6})$

Environment



$$\chi(\zeta) = \begin{bmatrix} N_1 & n_1^1 & m_1^1 & n_1^2 & m_1^2 & n_1^3 & m_1^3 \\ N_2 & n_2^1 & m_2^1 & n_2^2 & m_2^2 & n_2^3 & m_2^3 \\ N_3 & n_3^1 & m_3^1 & n_3^2 & m_3^2 & n_3^3 & m_3^3 \\ & & & \vdots & & & \\ N_k & n_k^1 & m_k^1 & n_k^2 & m_k^2 & n_k^3 & m_k^3 \end{bmatrix}_{k\times 7}$$

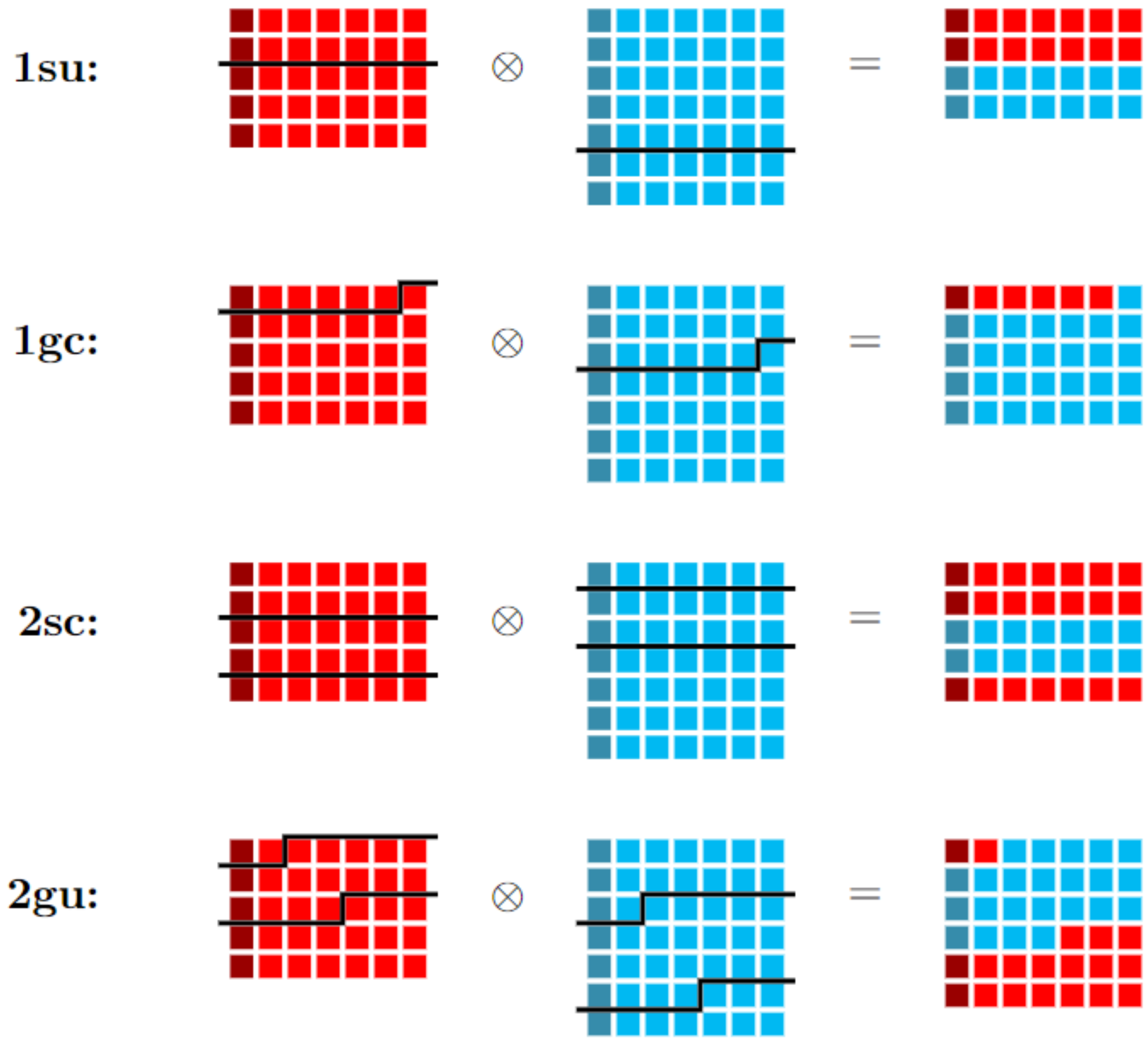
- Fix $b_i \in \{0,\frac{1}{2}\}$, $\widehat{U}_I > 0$, N_{max}, k_{min}, k_{max}.
- Winding numbers pairwise co-prime.
- **Env:** Possible restrictions of search space, e.g., replacement of $A \to A'$, $B \to B'$, $C \to C'$ branes or removing C, C' branes from the chromosomes and adding them later for maximizing fitness.

Crossover

- We employ 8 uniform cross-over methods:
  - **# crossover points (1/2)**
  - **crossovers between stacks/genes (s/g)**
  - **# stacks constrained to match with that of one of the parents (c/u)**

  [Loges, GS, ’21]

- Other cross-over methods can be employed.



Mutations

- **Change stack size:** $\mu_N(N) = N \pm 1$

- **Change winding number:** $\mu_w(w) = w \pm \{1,2\}$

- **Randomize winding number signs:**

    - e.g. $\mu_{\pm}(\{w\}) = (n^1, -m^1, n^2, m^2, -n^3, -m^3)$ flips the signs of $\widehat{X}^0, \widehat{X}^1, \hat{Y}^2, \hat{Y}^3$.

- **Permute $\widehat{X}^I, \hat{Y}^I$:**

    - e.g. $\mu_{\mathfrak{S}_4}(\{w\}) = (n^2, m^2, n^1, m^1, n^3, m^3)$ swaps $\widehat{X}^1 \leftrightarrow \widehat{X}^2$ and $\hat{Y}^1 \leftrightarrow \hat{Y}^2$.

Fitness

- The fitness function measures how close an individual is to having desirable properties.

- We take the fitness function to take the form ($\mathscr{F} = 0$ is optimal):

$$\mathcal{F}(\zeta) = \mathcal{W}_{\mathrm{T}}\mathcal{F}_{\mathrm{T}}(\zeta) + \mathcal{W}_{\mathrm{K}}\mathcal{F}_{\mathrm{K}}(\zeta) + \mathcal{W}_{\mathrm{S}}\mathcal{F}_{\mathrm{S}}(\zeta) + \mathcal{W}_{\mathrm{M}}\mathcal{F}_{\mathrm{M}}(\zeta)$$

where

$$\mathcal{F}_{\mathrm{T}}(\zeta) = h\Big(\tfrac{\langle T^I \rangle}{\Delta_{\mathrm{T}}}\Big)\,, \qquad T^I = \Big|\sum_a N_a \widehat{X}^I_a - 8\Big|\,, \qquad h(z) = \tfrac{z}{1+z}$$

$$\mathcal{F}_{\mathrm{K}}(\zeta) = \langle K^I \rangle\,, \qquad K^I = \Big(\sum_a N_a \widehat{Y}^I_a\Big) \mod 2\,,$$

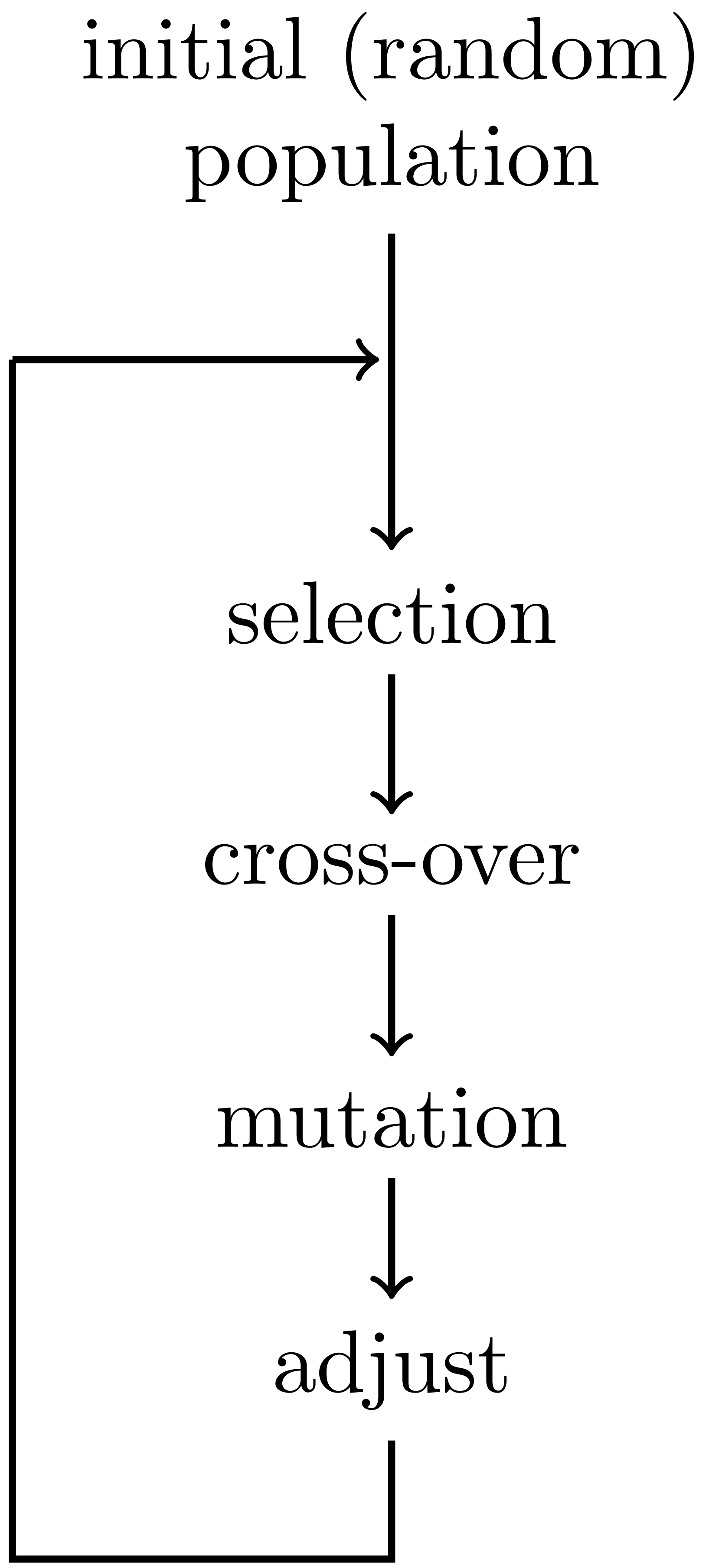
$$\mathcal{F}_{\mathrm{S}}(\zeta) = h\Big(\tfrac{\langle S_a \rangle}{\Delta_{\mathrm{S}}}\Big)\,, \qquad S_a = \left|\min\left\{\frac{\sum_I \widehat{X}^I_a \widehat{U}_I}{\sum_I \widehat{U}_I}, 0\right\}\right| + \left|\frac{\sum_I \widehat{Y}^I_a / \widehat{U}_I}{\sum_I 1/\widehat{U}_I}\right|\,,$$

$$\mathcal{F}_{\mathrm{M}}(\zeta) = G(\zeta)/4\,,$$

$G =$ "distance to the SM"

Brane Breeding: Summary

- **Adjustments:**
  - Non-empty stacks: each stack should have $N_a \geq 1$
  - Coprimality: $\gcd(n_a^i, m_a^i) = 1$
  - Standardization: bring $\bigotimes_{i=1}^{3} \left(n_a^i, m_a^i\right)$ to standard form
  - No-duplicates: stacks with identical $\bigotimes_{i=1}^{3} \left(n_a^i, m_a^i\right)$ are combined
- **Hyperparameter optimization:**
  - Environment parameters: $b_i,\ \widehat{U}_I, k_{\min}, k_{\max}, N_{\max}, \mathbf{env}$
  - Metaparameters: $n_{\text{pop}}, n_{\text{elite}}, g_{\max}$
  - Hyperparameters: $p_i, r_i, \mathscr{W}_i, \Delta_{\text{T,S}}$



initial (random) population

selection

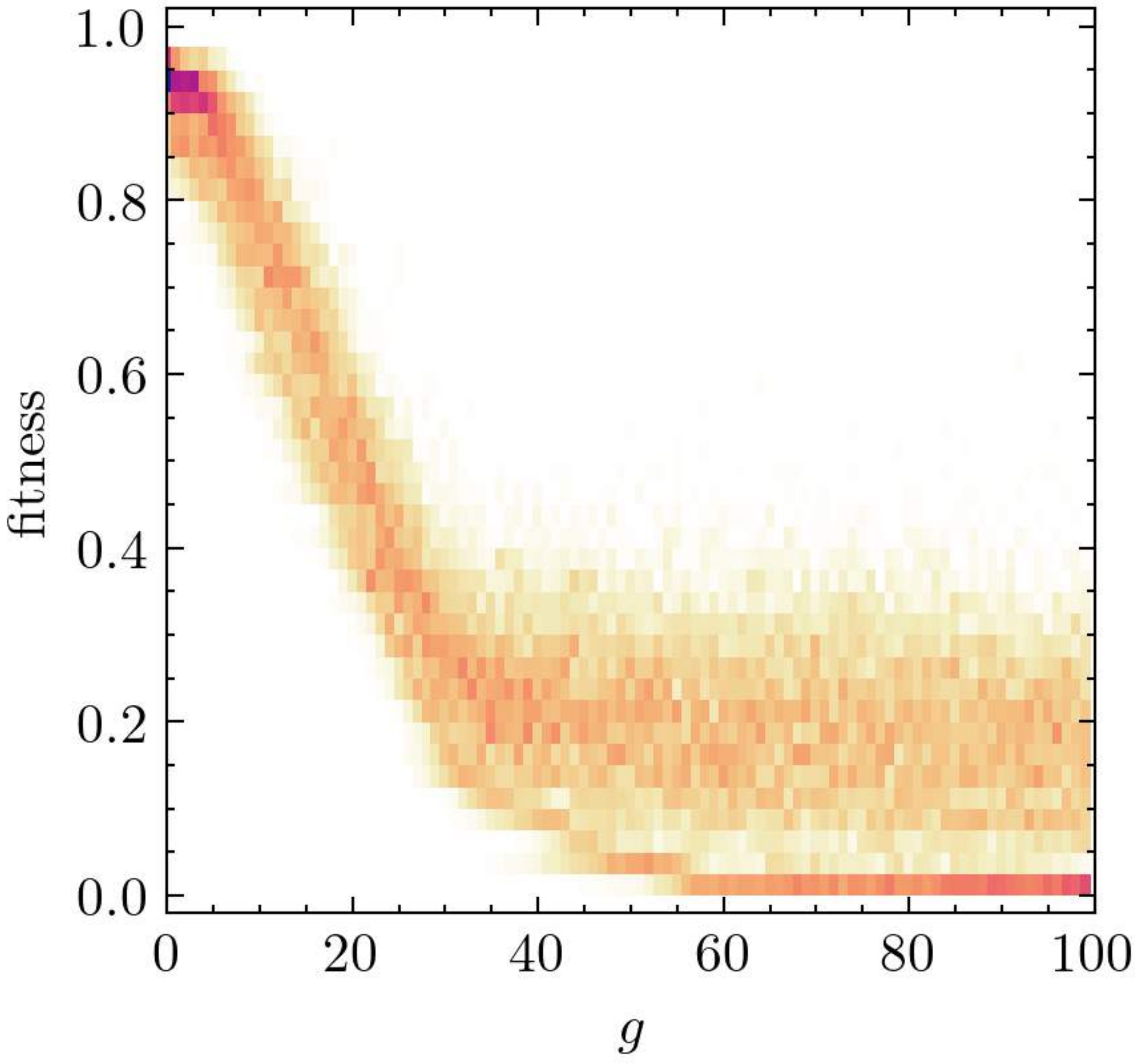
cross-over

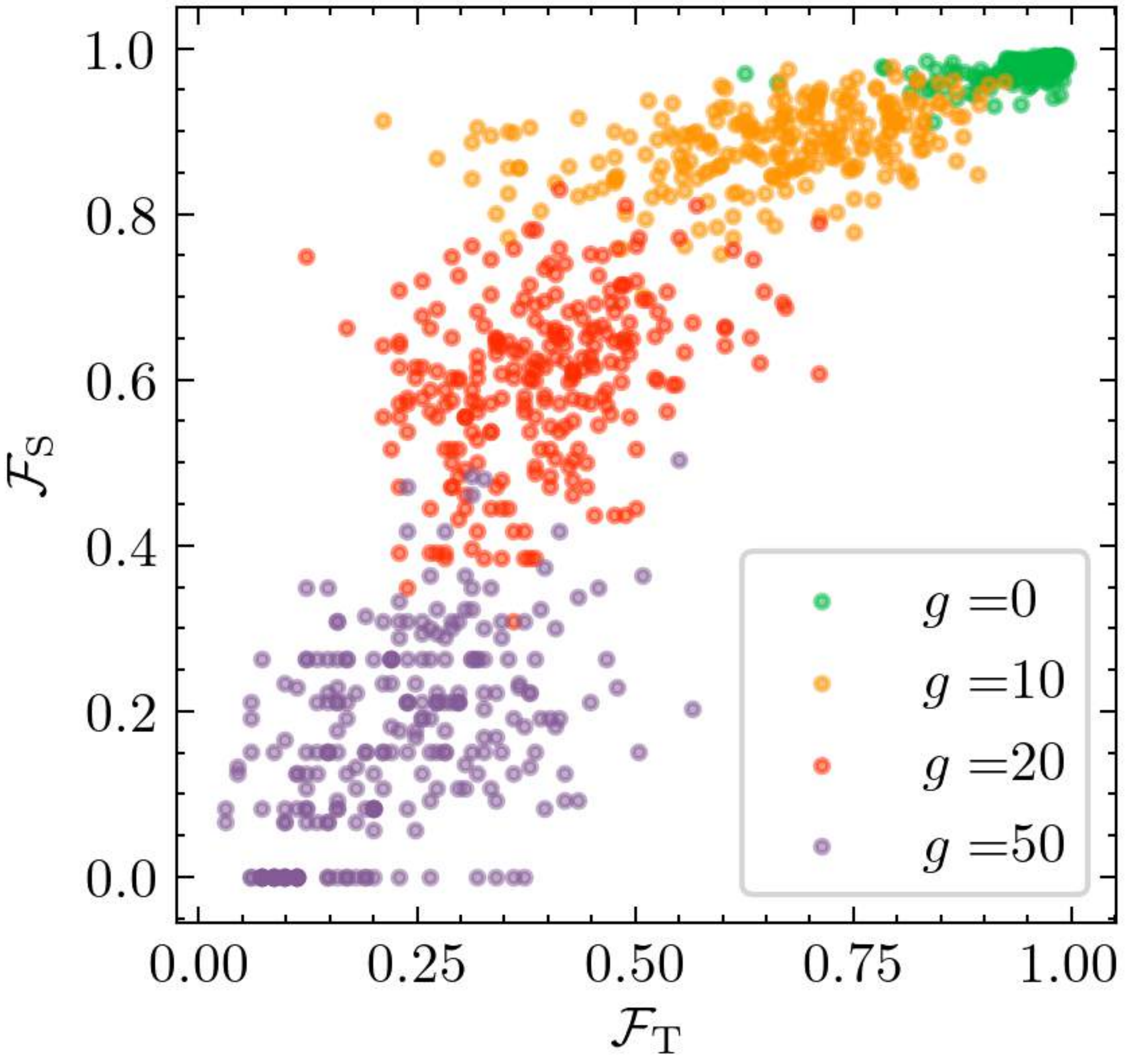
mutation

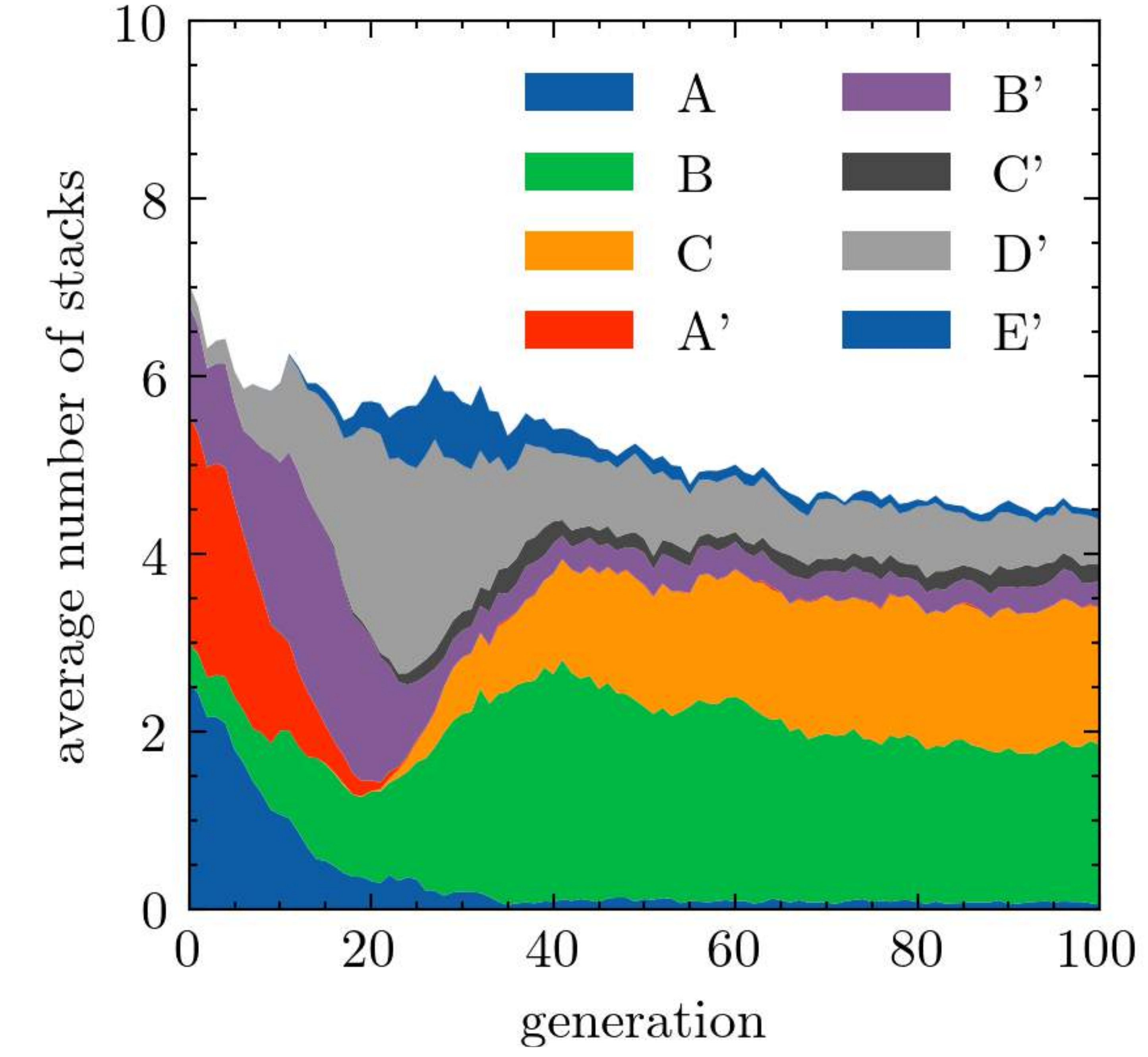
adjust

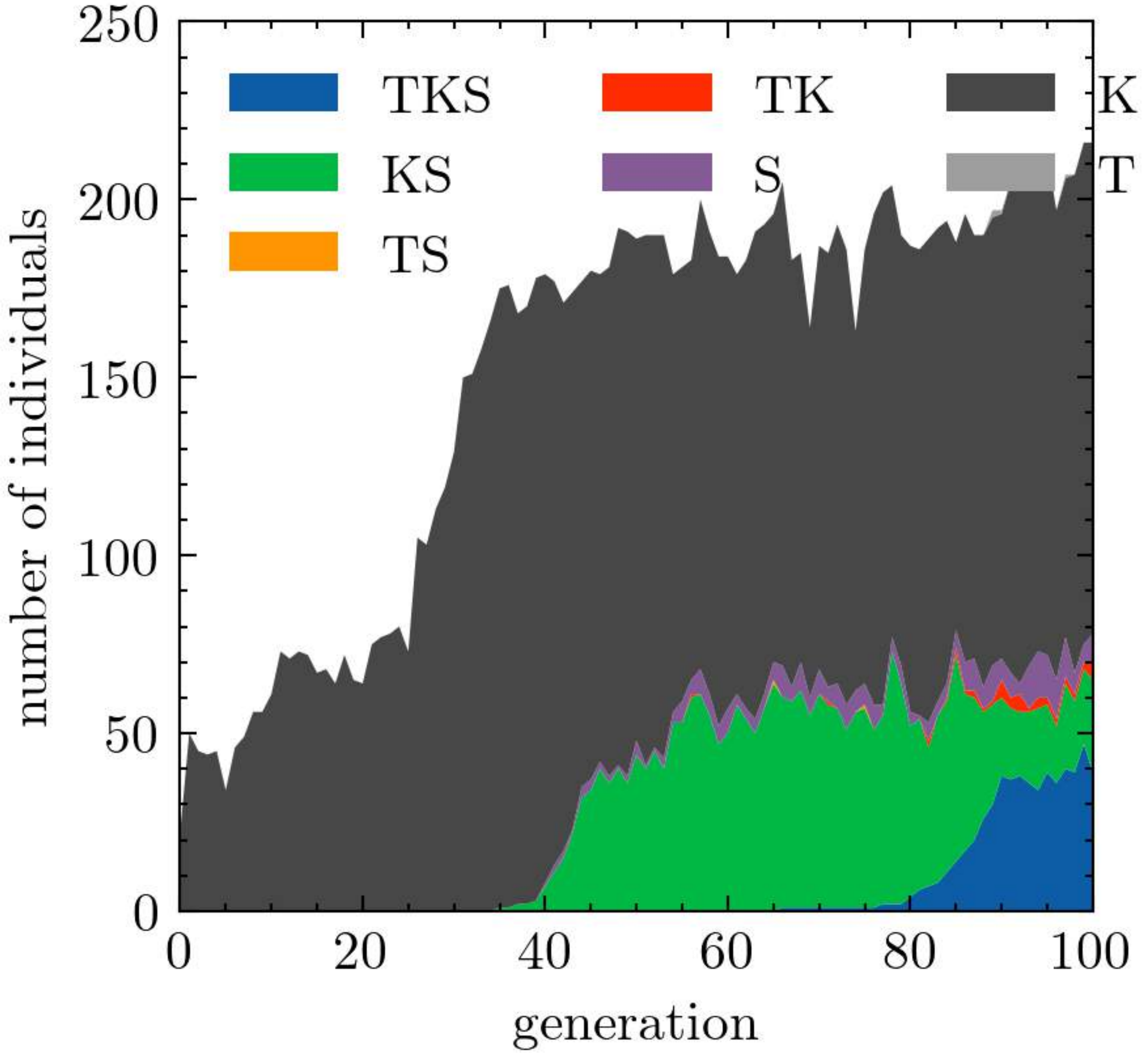
How GA Learns

$$b_i = 0, \quad \widehat{U}_I = 1, \quad (n_{\mathrm{pop}}, n_{\mathrm{elite}}, g_{\max}) = (250, 25, 100), \quad \mathcal{W}_{\mathrm{MSSM}} = 0$$



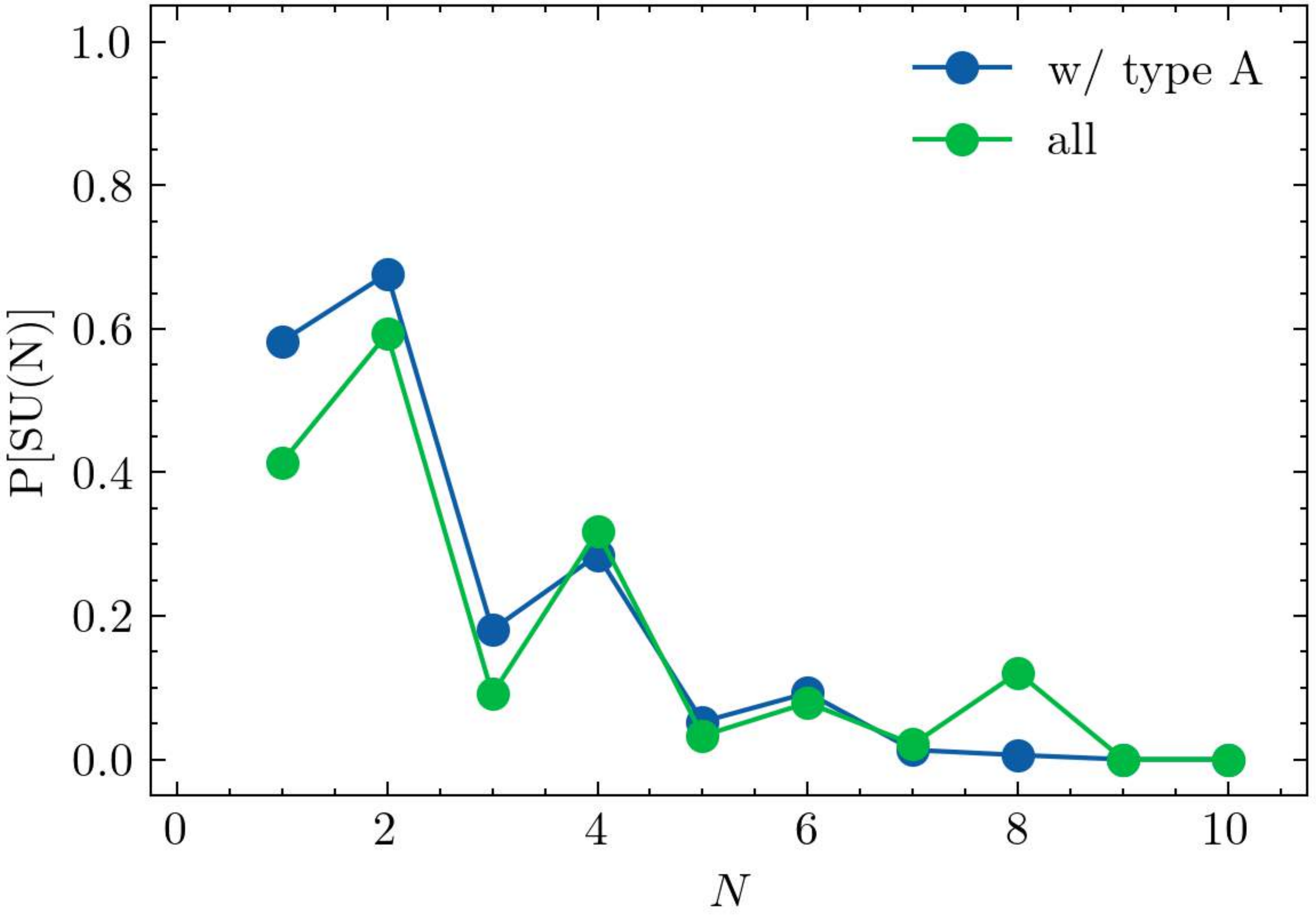


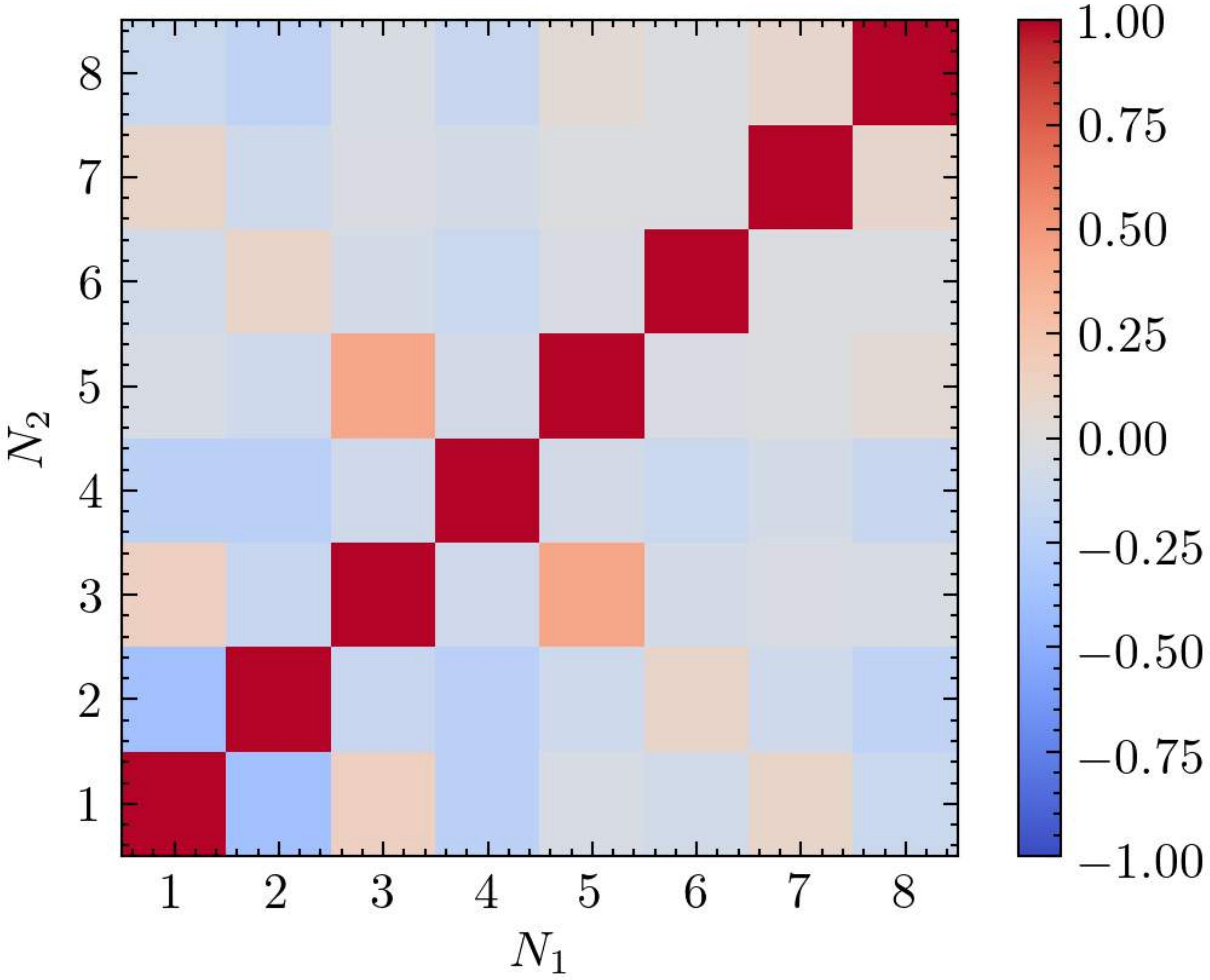




Landscape Statistics: U(N) Factors

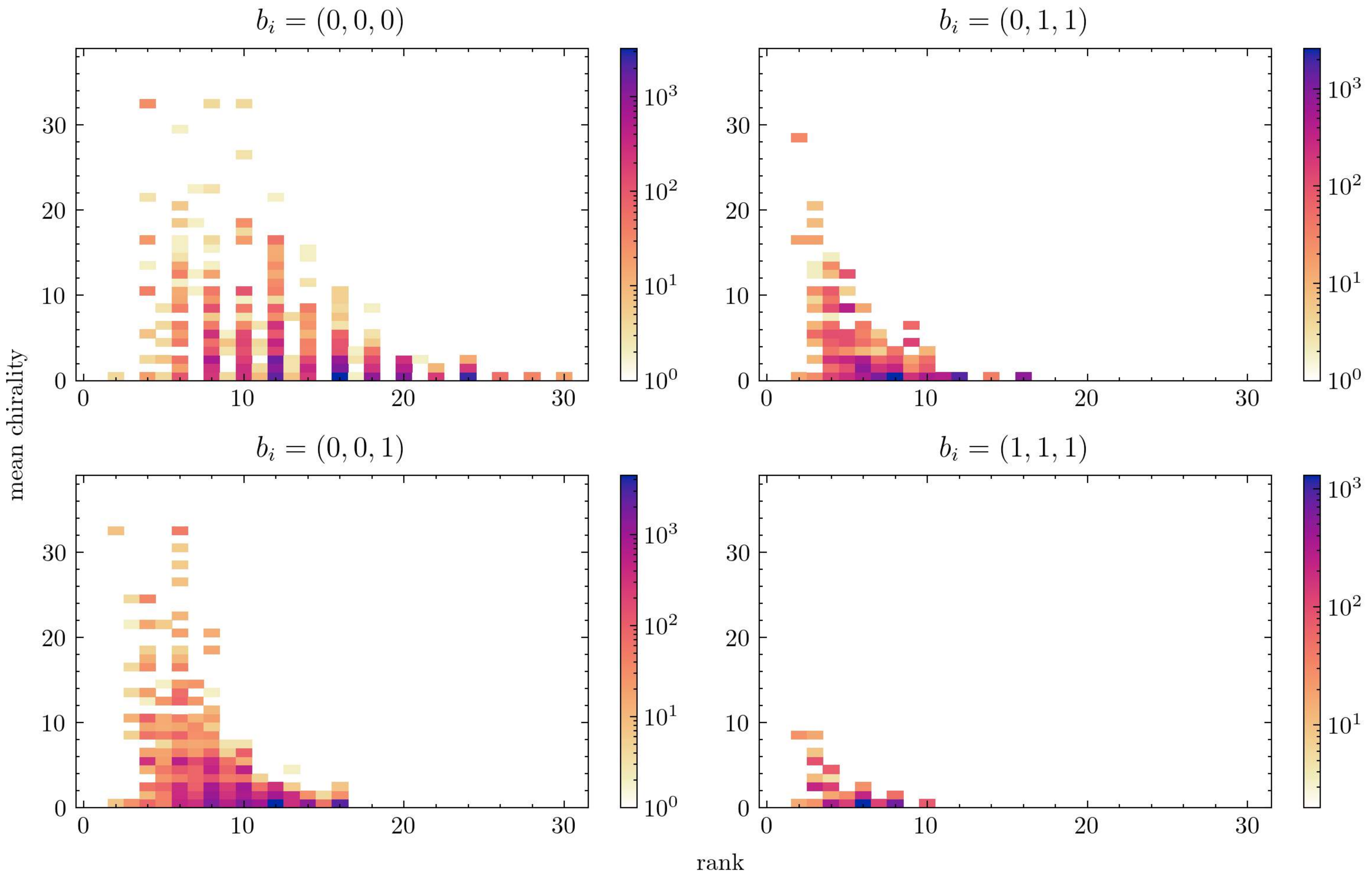
- $\mathcal{O}(10^6)$ unique, fully consistent models can be generated in $\mathcal{O}(\text{hours})$.

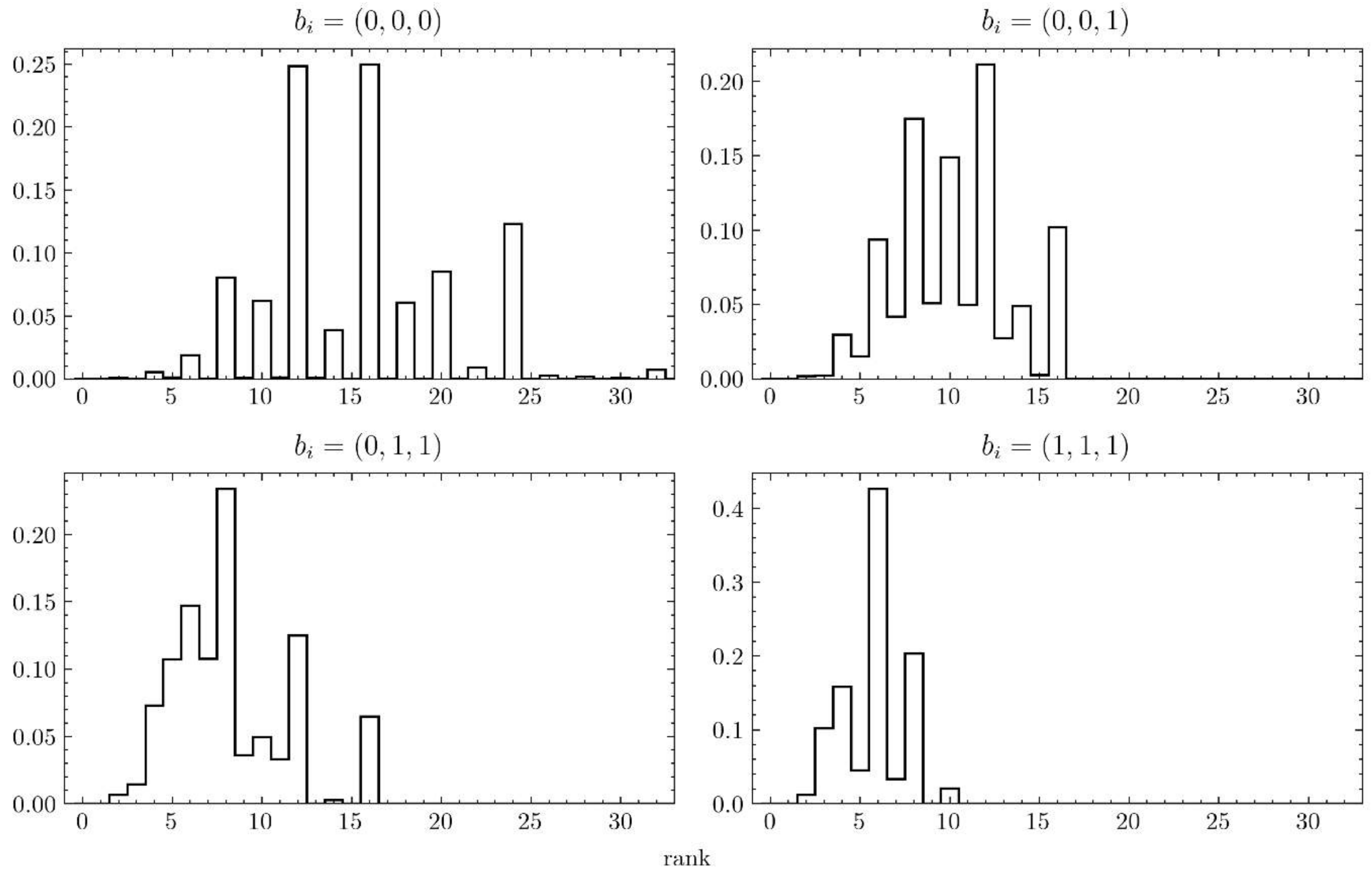




P[SU(2)] ≈ 0.5, P[SU(3)] ≈ 0.1, P[SU(2) and SU(3)] $\approx 0.04 \approx$ P[SU(2)] × P[SU(3)]

Landscape Statistics: Chirality vs Rank





Other correlations i) between environment with phenotypes (e.g. # generations, rank, etc), and ii) among phenotypes, plus iii) other landscape statistics can be found in [Loges, GS, '21].

Summary

- We demonstrated how GAs and RL can effectively find desirable string vacua: **flux vacua** [Cole, Schachner, GS, ’19];[Cole, Schachner, GS, ’21] and **intersecting brane models** [Loges, GS, ’21] .

- These methods **discover structure** (e.g., symmetries, correlations) in the landscape (resonate with [Cole, GS, ’18]) but in a **complementary** way: combining them can reduce sampling bias.

- Our studies highlighted the similarities/differences of the **optimization/learning strategies** used (e.g., both exploit the strategy of SUSY → filler branes but different vacua are sampled).

- **Preliminary landscape statistics** (stay tuned for our upcoming paper [Loges, GS, ’21]).

Summary

- We demonstrated how GAs and RL can effectively find desirable string vacua: **flux vacua** [Cole, Schachner, GS, '19];[Cole, Schachner, GS, '21] and **intersecting brane models** [Loges, GS, '21] .
- These methods **discover structure** (e.g., symmetries, correlations) in the landscape (resonate with [Cole, GS, '18]) but in a **complementary** way: combining them can reduce sampling bias.
- Our studies highlighted the similarities/differences of the **optimization/learning strategies** used (e.g., both exploit the strategy of SUSY → filler branes but different vacua are sampled).
- **Preliminary landscape statistics** (stay tuned for our upcoming paper [Loges, GS, '21]).

+

=