Genetic algorithms as a search tool for strings

SAA+J.Rizos, JHEP 1408 (2014) 010,1404.7359 hep-th

SAA+D.Cerdeno,S.Robles, 1805.03615 hep-ph

Overview

- String theories typically produce vast theory spaces.
- We would like to be able to find the “Standard Model” in them (or at least to check if a SM is there). We would like to find slightly AdS vacua.
- Such tasks are typically NP complete (difficulty increases exponentially with the search criteria, but the solution can be verified in polynomial time).
- Heuristic search techniques are effective in such problems. Here I will discuss genetic algorithms - based on evolutionary dynamics.
- The string theory example I will consider is in the Free-Fermionic formulation but the same techniques could be applied to many constructions.
- Using the pMSSM as a toy, I wish to show how GAs can be used to probe a parameter space. (There is no statistical data but there is a picture of the structure of the “fitness” landscape.)

GA work in particle theory …

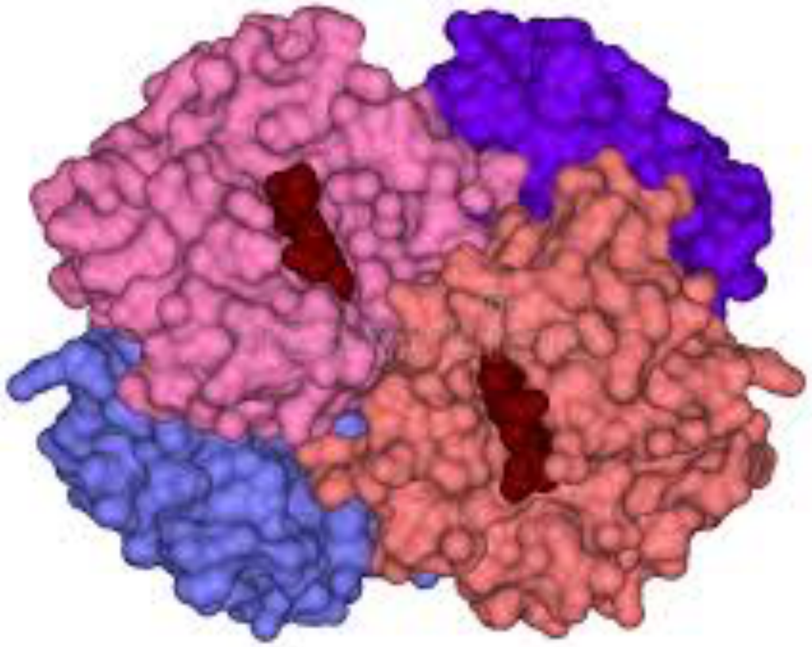
- Yamaguchi and H. Nakajima (2000)
- B. C. Allanach, D. Grellscheid and F. Quevedo (2004)
- Y. Akrami, P. Scott, J. Edsjo, J. Conrad and L. Bergstrom (2009)
- J. Bl˚aba¨ck, U. Danielsson and G. Dibitetto, (2013)

On the largeness (or otherwise) of 10^{500}

- Consider biological landscapes: problems that were solved by evolution

e.g. Haemoglobin molecule.

$C_{2932}H_{4724}N_{828}O_{840}S_8Fe_4$



- 2 legs of 141 amino acids, plus 2 legs of 146. 20 amino acids means …10^{747} !!
- Or possibly we should estimate #choices of C,H,…Fe from 92 elements ..10^{18334} !!!

Example of dealing with a string sized landscape

- GA’s (based on evolutionary dynamics) work most effectively when

(Holland, E.David, Reeves+Rowe, Jones+Forrest)

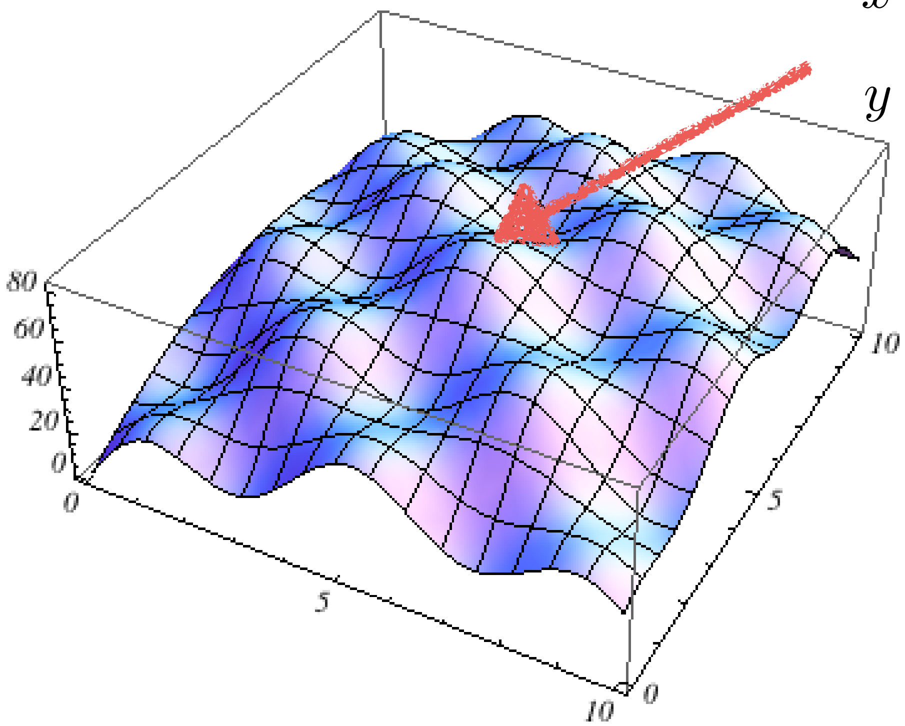
a) many criteria being applied at the same time

b) good correlation between “goodness of fit” and “closeness to maximum” (Fitness/Distance Correlation)

Disadvantage: by their nature statistical information very hard/impossible to get

Example: find maximum point to accuracy of 250 decimal places without using calculus.

$$\begin{aligned} x &= a.bcdef\ldots \\ y &= g.hijkl\ldots \end{aligned} \quad \Longrightarrow \quad 10^{500}$$



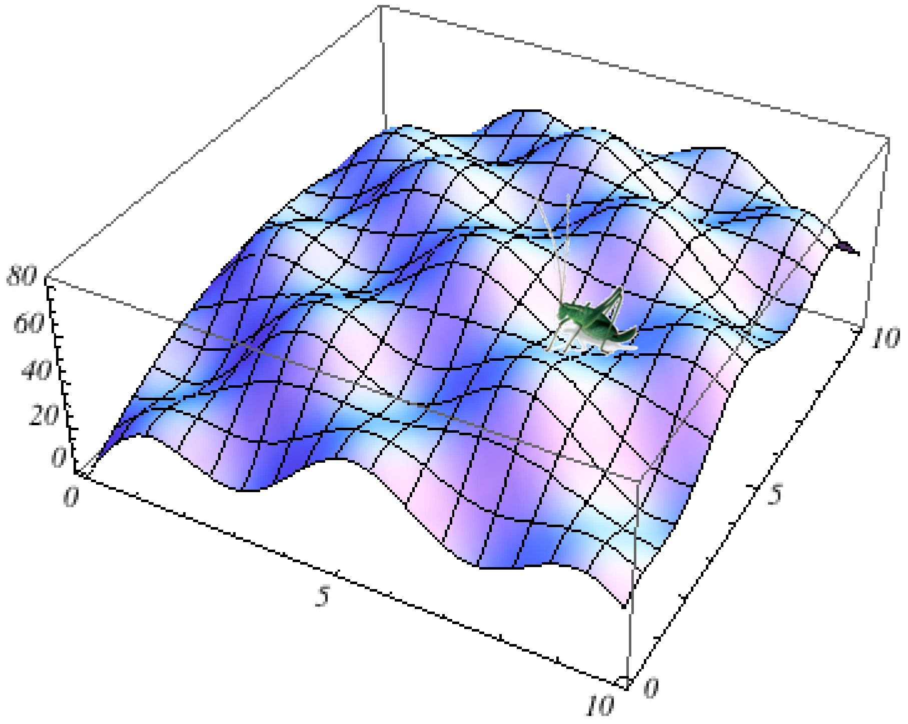
$$f(x,y) = 12\left(\cos\frac{3y}{2}\,\sin\frac{3x}{2} + x + y\right) - x^2 - y^2.$$

Example of dealing with a string sized landscape

- Define a “creature” and write out its coordinates => genotype
- Terminology: Genotype = data. Phenotype = *f(x,y)*.

$$x = a.bcdef...$$

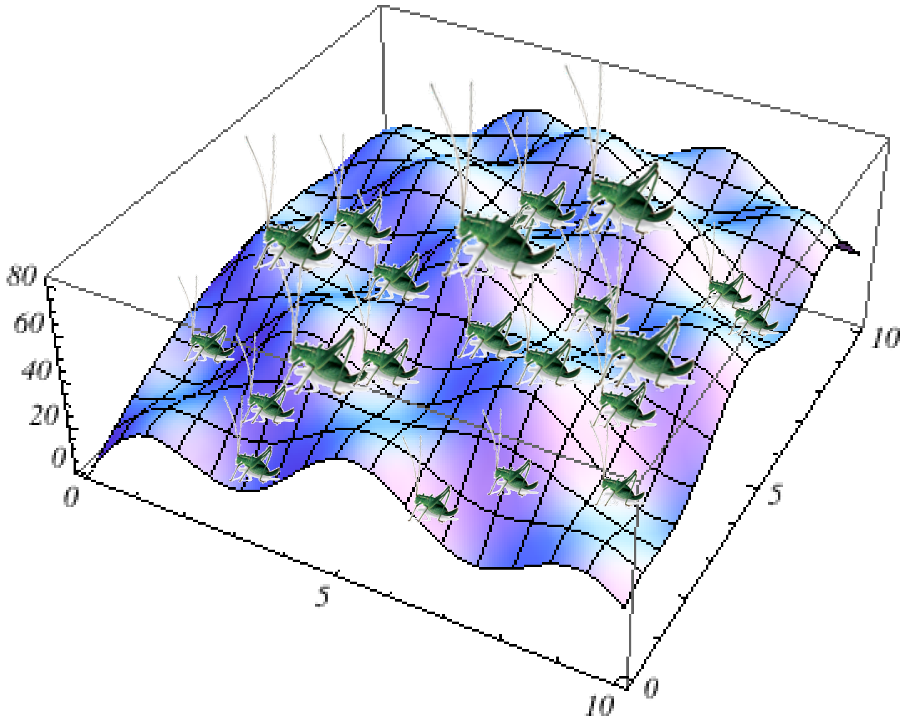
$$y = g.hijkl...$$



Example of dealing with a string sized landscape

- Population initially sprinkled at random

- Step1: Define fitness function, *f(x,y)*. Selection for breeding will be based on **fitness** (e.g. *f* = height in this case).

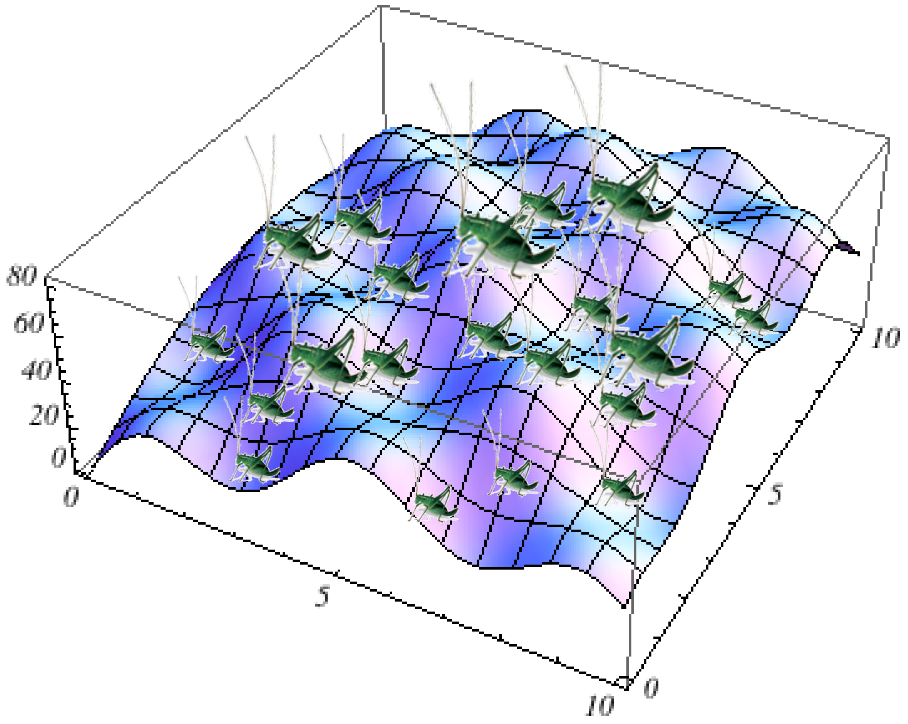


Example of dealing with a string sized landscape

- Population initially sprinkled at random

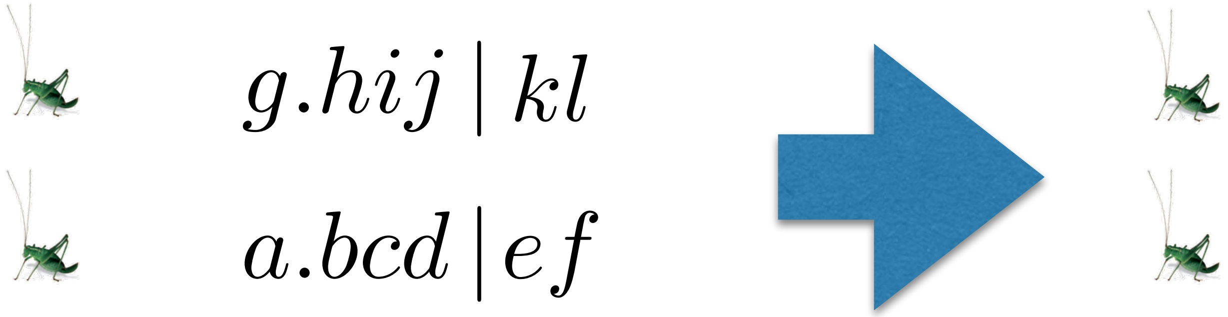
- Step2: **Selection**. Select pairs for breeding such that the most fit individuals can breed several times, while unfit ones might not breed at all: e.g. “roulette wheel”.

$$p_i = \frac{1}{p}\frac{(\alpha - 1)\left(f_i - \bar{f}\right) + \left(f_{max} - \bar{f}\right)}{f_{max} - \bar{f}},$$



Simple example of a string sized landscape

- Step 3: **breeding**: cut and splice genotypes of breeding pairs somehow (not really crucial how)



$$g.hij \,|\, kl$$

$$a.bcd \,|\, ef$$

Simple example of a string sized landscape

- Step 4: **Mutation** of a randomly chosen small percentage of digits (alleles).

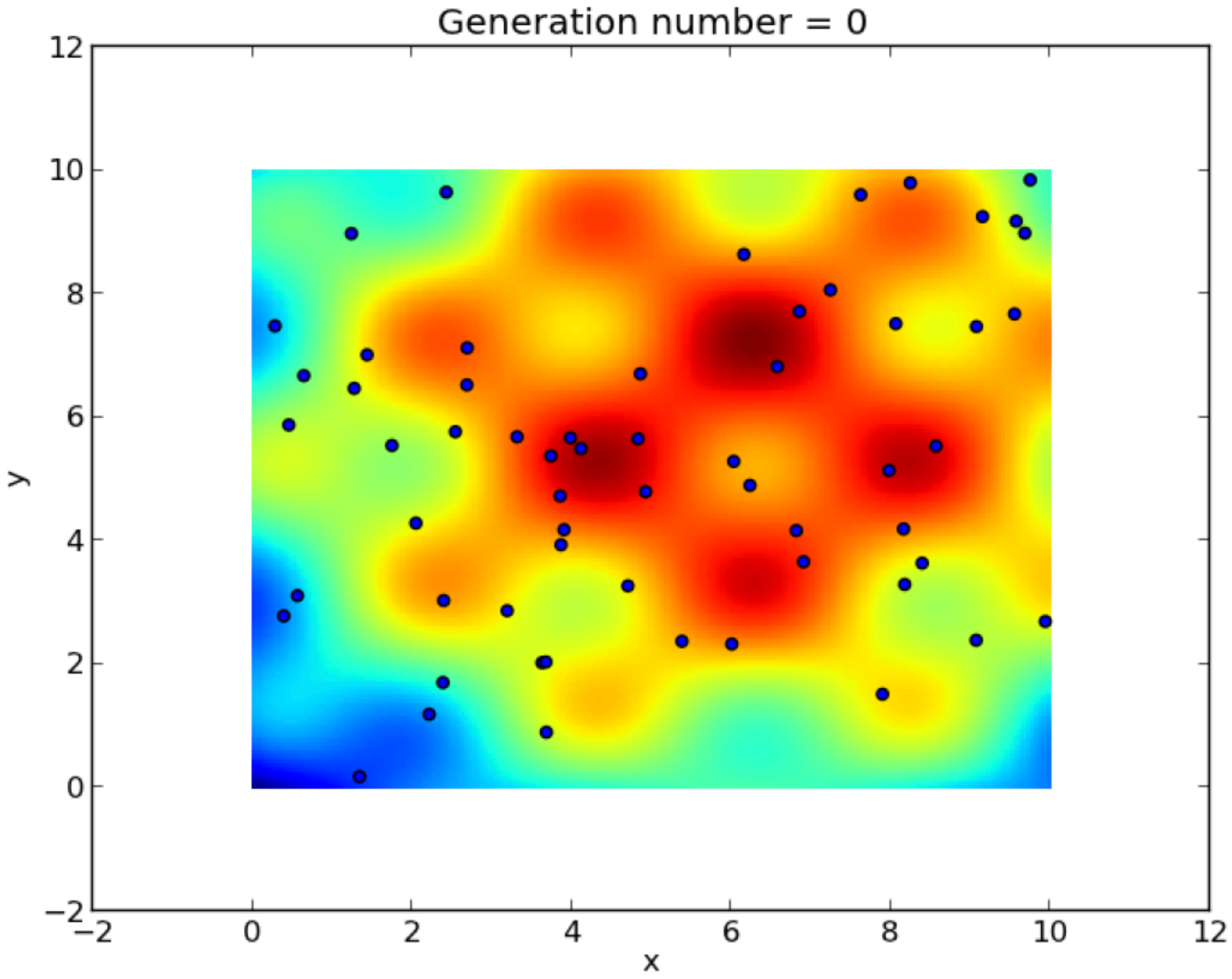
$$a.bcdefghij\ldots$$

$$a.bcdef'ghi'j\ldots$$

- Steps 5 … infinity: rinse and repeat

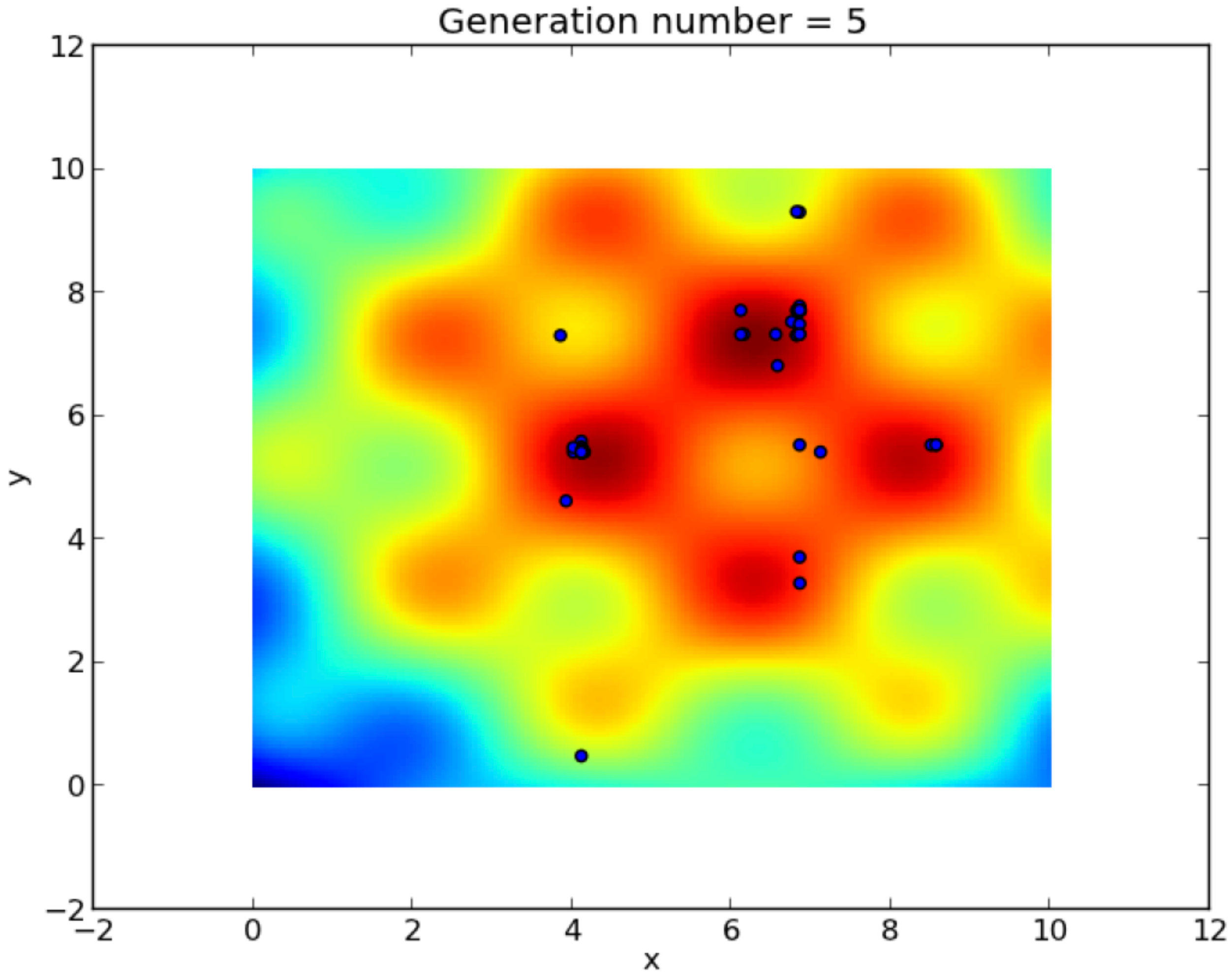
Simple example of a string sized landscape

- Summary: three crucial ingredients ***Selection*** (favours the optimisation); ***Breeding/crossover*** (propagates favourable “schema” - Holland); ***Mutation*** (prevents stagnation: evolution proceeds by punctuated equilibria)



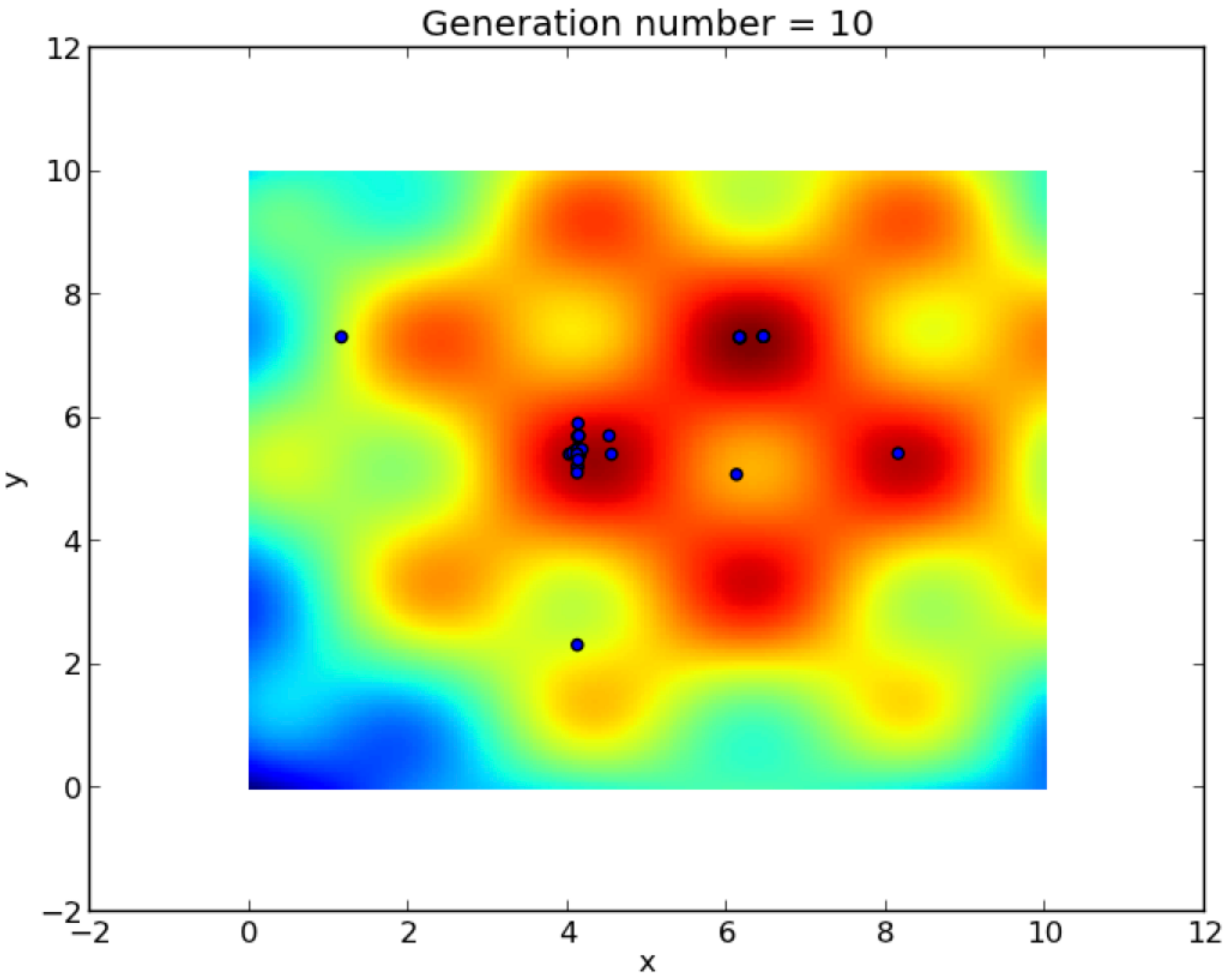
Simple example of a string sized landscape

- Summary: three crucial ingredients ***Selection*** (favours the optimisation); ***Breeding/crossover*** (propagates favourable “schema” - Holland); ***Mutation*** (prevents stagnation: evolution proceeds by punctuated equilibria)



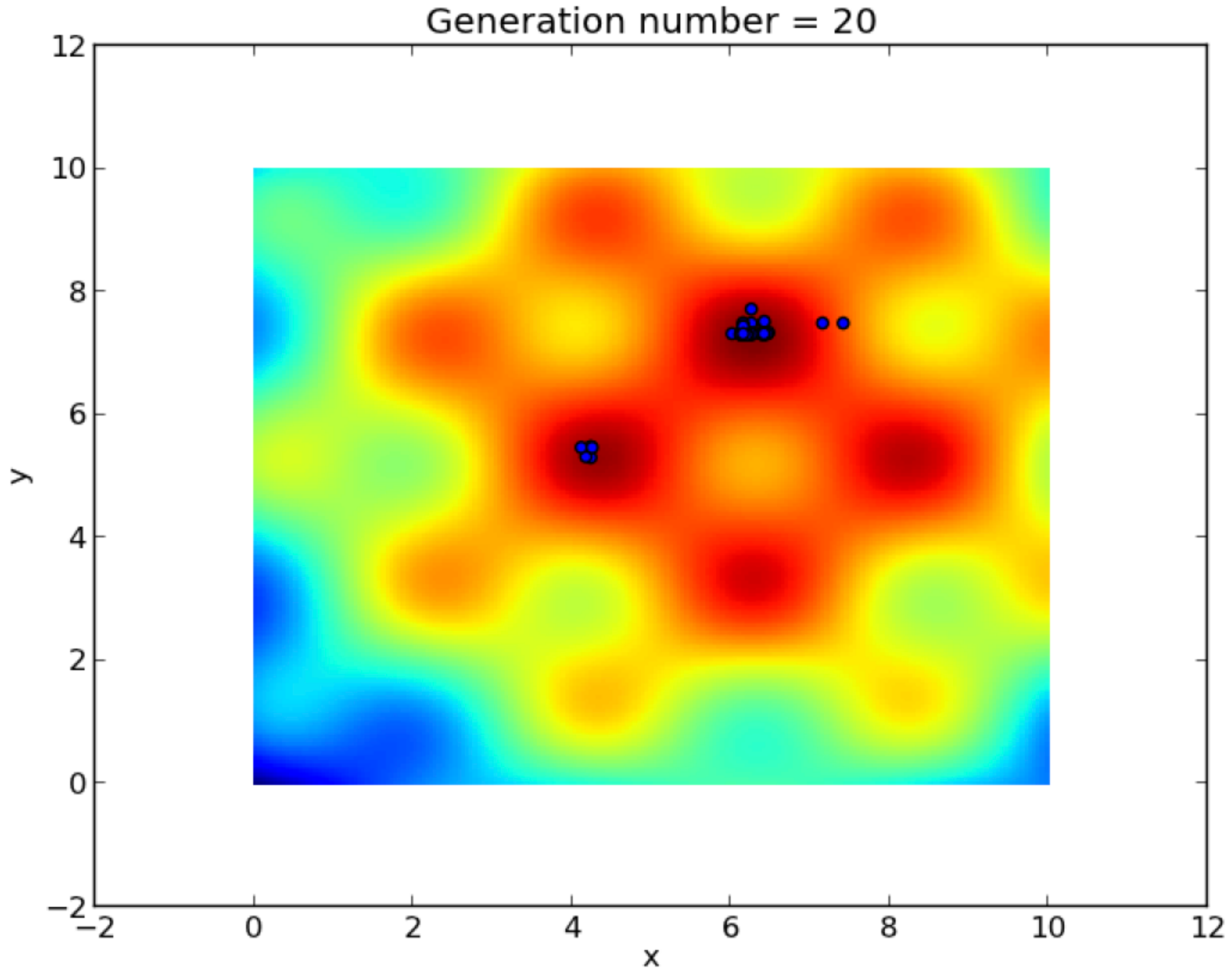
Simple example of a string sized landscape

- Summary: three crucial ingredients ***Selection*** (favours the optimisation); ***Breeding/crossover*** (propagates favourable "schema" - Holland); ***Mutation*** (prevents stagnation: evolution proceeds by punctuated equilibria)



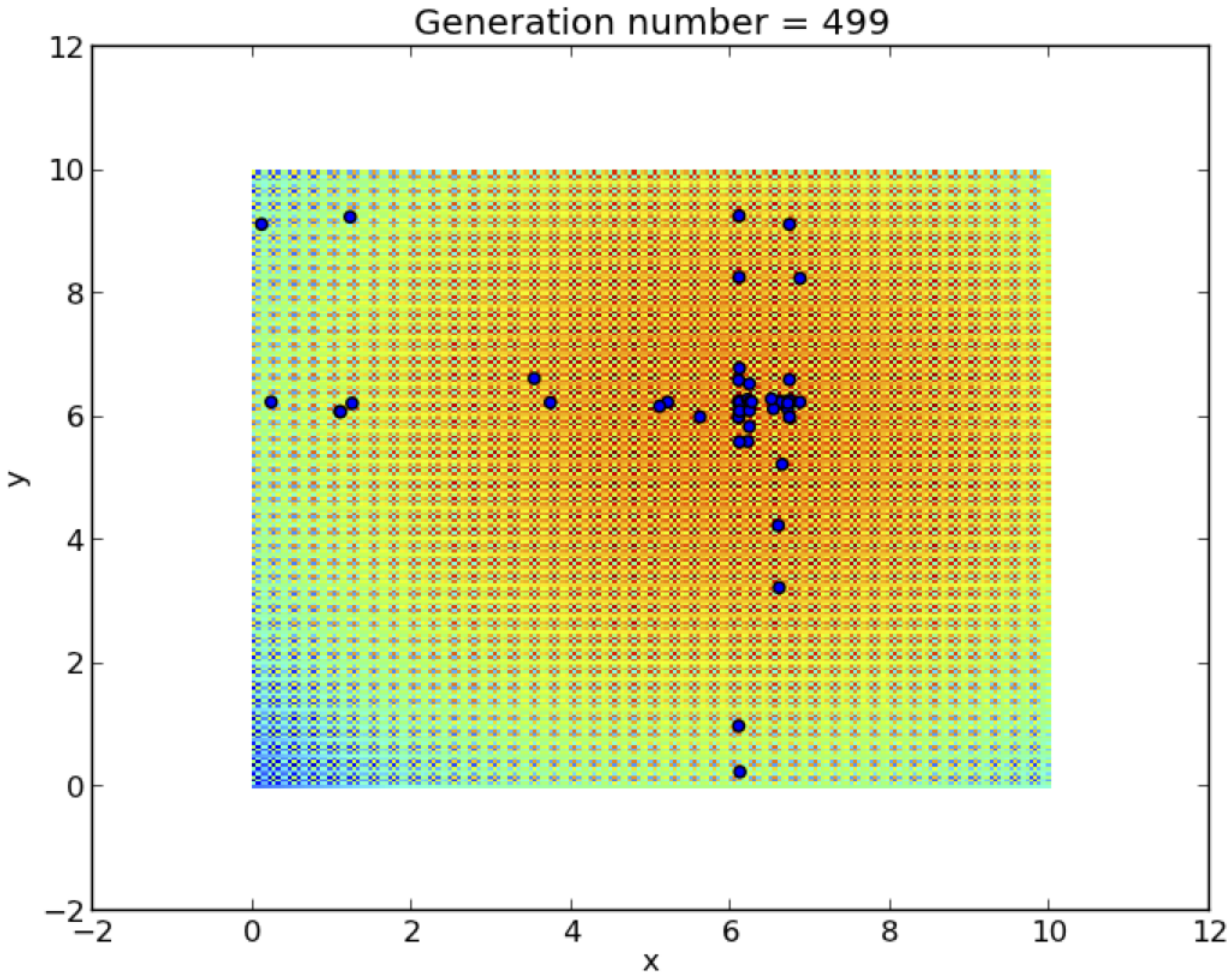
Simple example of a string sized landscape

- Summary: three crucial ingredients ***Selection*** (favours the optimisation); ***Breeding/crossover*** (propagates favourable “schema” - Holland); ***Mutation*** (prevents stagnation: evolution proceeds by punctuated equilibria)



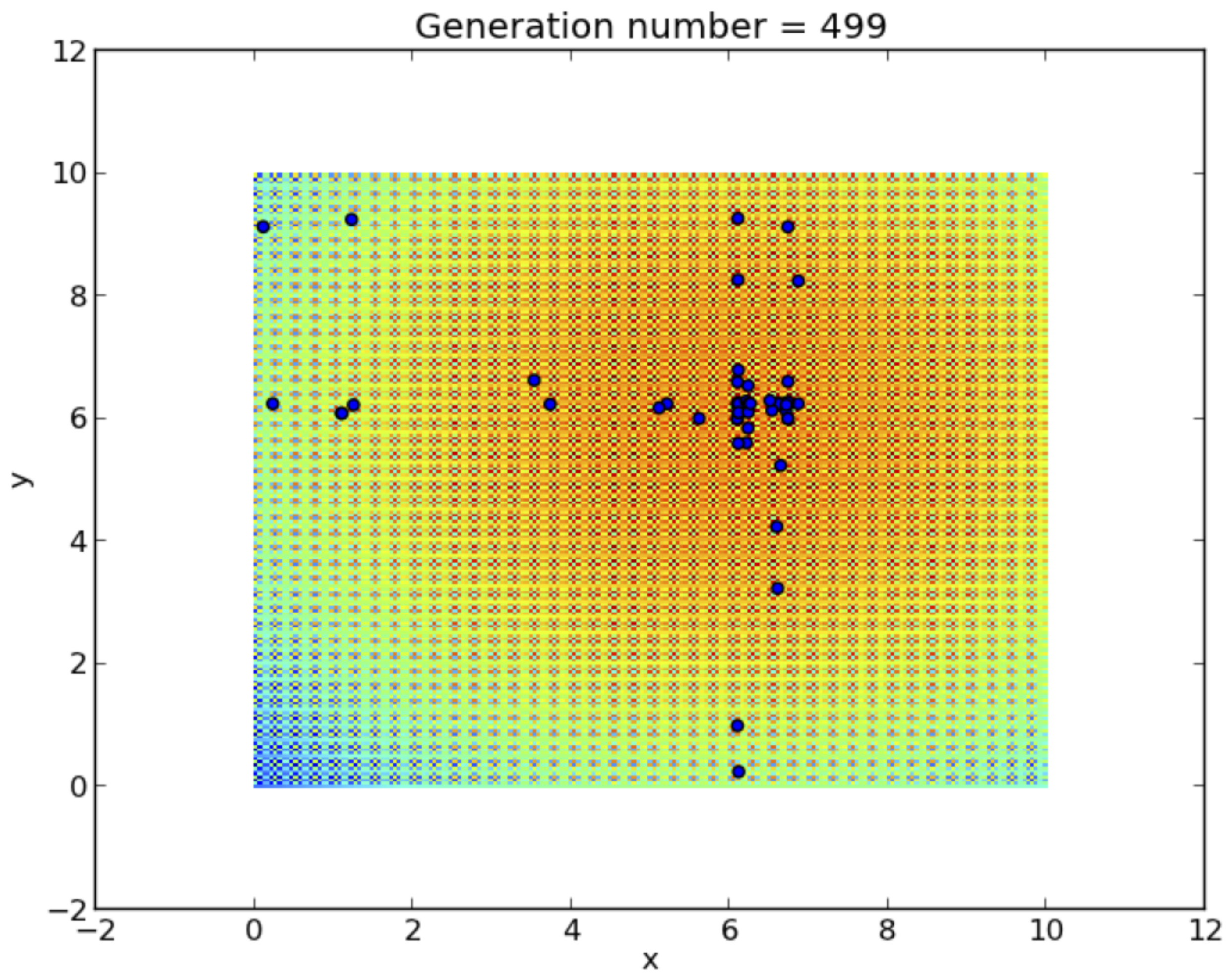
Simple example of a string sized landscape

- Summary: three crucial ingredients ***Selection*** (favours the optimisation); ***Breeding/crossover*** (propagates favourable “schema” - Holland); ***Mutation*** (prevents stagnation: evolution proceeds by punctuated equilibria)



Simple example of a string sized landscape

- Warning: in this example the convergence to a solution is easy to visualise: in strings it is very hard (high dimensionality - later)
- NB: in general the optimisation function does not have to be continuous or differentiable.

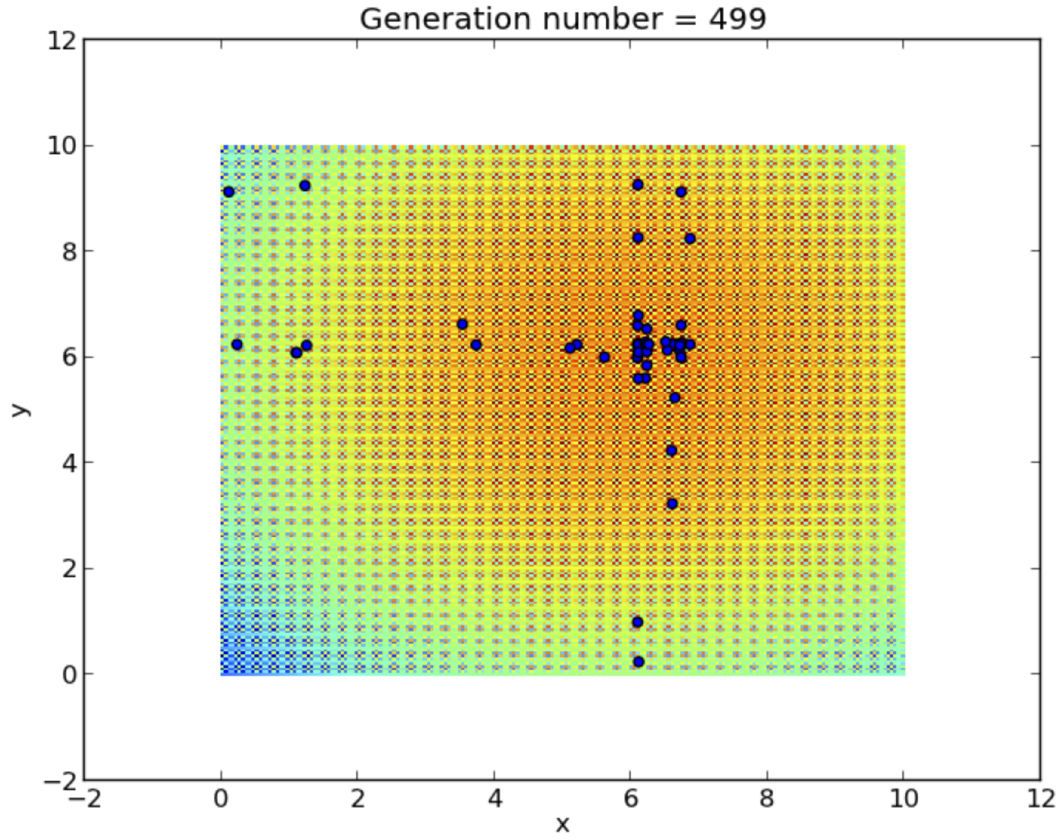


Schemata $S = 3{*}{*}{*}4{*}6$

- Holland proposed a probabilistic explanation for the efficiency of genetic algorithms: suppose we have *n(S,t)* members of population with schema *S*
- With simple probabilistic arguments one can incorporate the effect of a single-point crossover destroying *S*, and mutations at a rate p_m per allele to find a lower bound

$$n(S,t+1) \geq n(S,t)\frac{f_S(t)}{\bar{f}}\left(1-\frac{d(S)}{l-1}\right)(1-p_m)^{o(S)}$$

avge fitness of members with *S*
order *o= 3*
defining length *d=7*



In this example the leading digits of x and y are schemata

Schemata $S = 3{*}{*}{*}4{*}6$

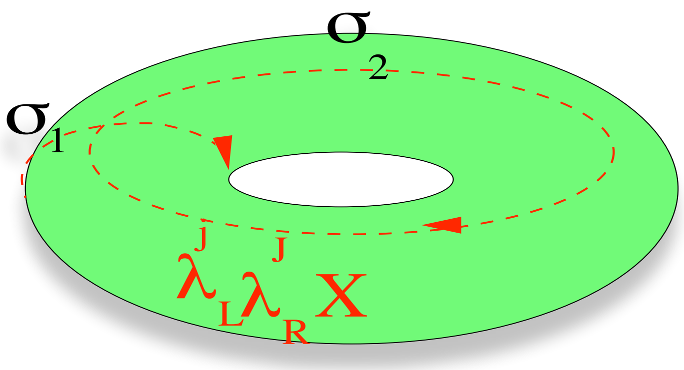
- Initial growth of *n(S,t)* is exponential
- At late times find equilibrium for average fitness determined by p_m
- Selection pushes towards convergence
- Mutation pushes system away from convergence

Optimisation:

- Like any machine learning technique you can run into problems unless you optimise …
  - Fitness — rank selection often works best to overcome flat maxima
  - Selection — Elitist selection (copy fittest individual into new population and kill weakest). Also tournament selection, roulette wheel, etc
  - Breeding — two or more point cross-over to avoid edge effects
  - Mutation: check this is optimised (See later)
  - Creep mutation to overcome “Hamming walls” e.g. 0.999… ~ 1.0000… :

Simple optimisation problem

- Find a phenomenologically attractive Pati-Salam model.
- We will consider the Free-Fermionic formulation. (We know the answer by the way - since we want to test our technique!).
- We'll use the "fermionic string construction". These are general 4D models in which the world sheet degrees of freedom are fermions. (Kawai, Lewellyn, Tye; Antoniadis, Bachas, Kounnas)
- ***A single W/S fermion acquires phases u,v going round the 2 cycles of the torus:***



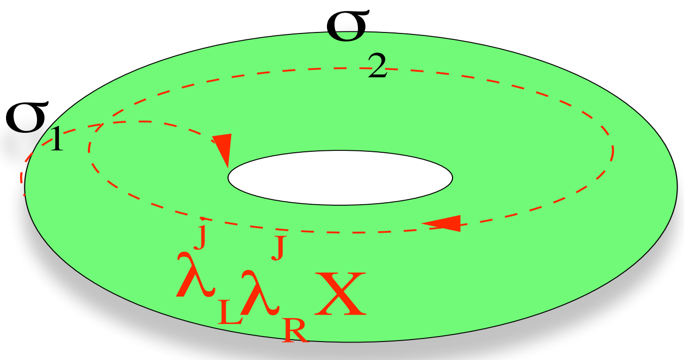
Simple optimisation problem

Models are defined in terms of a set of basis vectors and a set of phases associated with generalised GSO projections (GGSO).

$$\{v_1, v_2, \dots, v_{13}\} \qquad c\begin{bmatrix} v_i \\ v_j \end{bmatrix}, i, j = 1, \dots, n$$

we will use the following set: (Faraggi, Kounnas, Nooij, Rizos)

$$
\begin{aligned}
v_1 = \mathbb{1} &= \{\psi^\mu,\ \chi^{1,\dots,6}, y^{1,\dots,6}, \omega^{1,\dots,6} | \bar{y}^{1,\dots,6}, \bar{\omega}^{1,\dots,6}, \bar{\eta}^{1,2,3}, \bar{\psi}^{1,\dots,5}, \bar{\phi}^{1,\dots,8}\} \\
v_2 = S &= \{\psi^\mu, \chi^{1,\dots,6}\} \\
v_{2+i} = e_i &= \{y^i, \omega^i | \bar{y}^i, \bar{\omega}^i\},\ i = 1, \dots, 6 \\
v_9 = b_1 &= \{\chi^{34}, \chi^{56}, y^{34}, y^{56} | \bar{y}^{34}, \bar{y}^{56}, \bar{\eta}^1, \bar{\psi}^{1,\dots,5}\} \\
v_{10} = b_2 &= \{\chi^{12}, \chi^{56}, y^{12}, y^{56} | \bar{y}^{12}, \bar{y}^{56}, \bar{\eta}^2, \bar{\psi}^{1,\dots,5}\} \\
v_{11} = z_1 &= \{\bar{\phi}^{1,\dots,4}\} \\
v_{12} = z_2 &= \{\bar{\phi}^{5,\dots,8}\} \\
v_{13} = \alpha &= \{\bar{\psi}^{45}, \bar{y}^{1,2}\}\ .
\end{aligned}
$$



Simple optimisation problem

Our genotype will be the phases: $c\begin{bmatrix} v_i \\ v_j \end{bmatrix}, i, j = 1, \ldots, n$

$$c_{ij} = \begin{array}{c|ccccccccccccc|}
 & \mathbb{1} & S & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & b_1 & b_2 & z_1 & z_2 & \alpha \\
\hline
\mathbb{1} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
S & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
e_1 & 1 & 1 & 0 & \ell_{26} & \ell_{27} & \ell_{28} & \ell_{29} & \ell_{30} & \ell_6 & 0 & \ell_{14} & \ell_{20} & \ell_{41} \\
e_2 & 1 & 1 & \ell_{26} & 0 & \ell_{31} & \ell_{32} & \ell_{33} & \ell_{34} & \ell_7 & 0 & \ell_{15} & \ell_{21} & \ell_{42} \\
e_3 & 1 & 1 & \ell_{27} & \ell_{31} & 0 & \ell_{35} & \ell_{36} & \ell_{37} & 0 & \ell_{10} & \ell_{16} & \ell_{22} & \ell_{43} \\
e_4 & 1 & 1 & \ell_{28} & \ell_{32} & \ell_{35} & 0 & \ell_{38} & \ell_{39} & 0 & \ell_{11} & \ell_{17} & \ell_{23} & \ell_{44} \\
e_5 & 1 & 1 & \ell_{29} & \ell_{33} & \ell_{36} & \ell_{38} & 0 & \ell_{40} & \ell_8 & \ell_{12} & \ell_{18} & \ell_{24} & \ell_{45} \\
e_6 & 1 & 1 & \ell_{30} & \ell_{34} & \ell_{37} & \ell_{39} & \ell_{40} & 0 & \ell_9 & \ell_{13} & \ell_{19} & \ell_{25} & \ell_{46} \\
b_1 & 0 & 0 & \ell_6 & \ell_7 & 0 & 0 & \ell_8 & \ell_9 & 1 & 0 & \ell_2 & \ell_4 & \ell_{47} \\
b_2 & 0 & 0 & 0 & 0 & \ell_{10} & \ell_{11} & \ell_{12} & \ell_{13} & 0 & 1 & \ell_3 & \ell_5 & \ell_{48} \\
z_1 & 1 & 1 & \ell_{14} & \ell_{15} & \ell_{16} & \ell_{17} & \ell_{18} & \ell_{19} & \ell_2 & \ell_3 & 1 & \ell_1 & \ell_{49} \\
z_2 & 1 & 1 & \ell_{20} & \ell_{21} & \ell_{22} & \ell_{23} & \ell_{24} & \ell_{25} & \ell_4 & \ell_5 & \ell_1 & 1 & \ell_{50} \\
\alpha & 1 & 1 & \ell_{41} & \ell_{42} & \ell_{43} & \ell_{44} & \ell_{45} & \ell_{46} & \ell_{47}+1 & \ell_{48}+1 & \ell_{49}+1 & \ell_{50} & \ell_{51}
\end{array} \mod 2$$

51 independent phases in these models: $2^{51} = 2 \times 10^{15}$

Simple optimisation problem

This search space is (just about) searchable deterministically so we can compare the two methods. (Assel, Christodoulides, Faraggi, Kounnas, Rizos)

The phases determine the characteristics of the models

(a) 3 complete family generations, $n_g = 3$

(b) Existence of PS breaking Higgs, $k_R \geq 1$

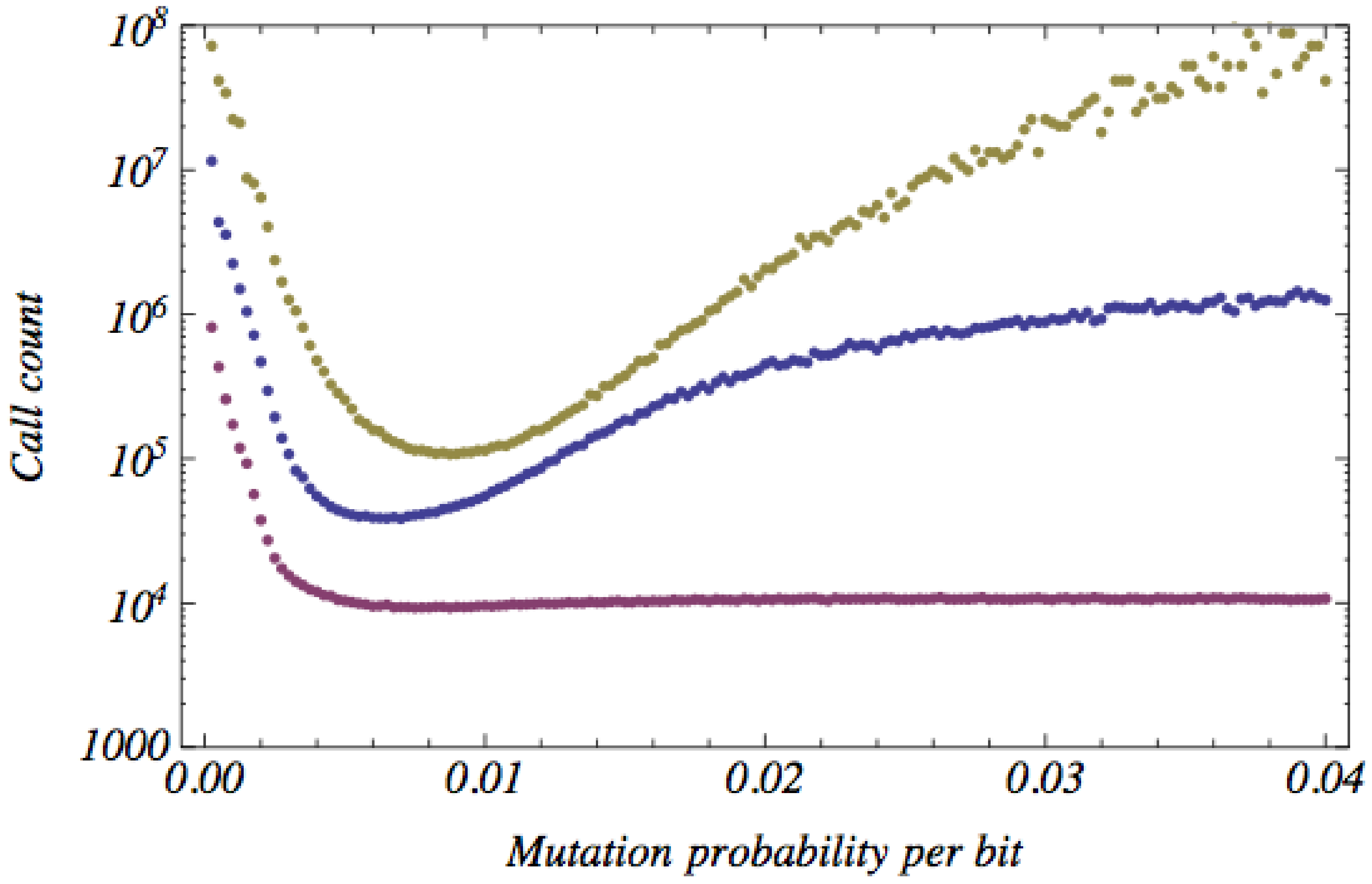
(c) Existence of SM Higgs doublets, $n_h \geq 1$

(d) Absence of exotic fractional charge states, $n_e = 0$

(e) Existence of top Yukawa coupling

- a)+b)+c) = 1 : 10,000
- a)+b)+c)+d) = 1 : 2,500,000
- a)+b)+c)+d)+e) = 1 : 10,000,000,000
- deterministically we would expect to have to construct 10 billion models to find an example of the latter

Simple optimisation problem



- Optimum mutation rate => genetic algorithm is working as expected
- GA's do not confer much advantage when the search is "easy"
- They work best when there are many criteria and the search is difficult =>

Fitness Distance Correlation

(Jones+Forrest;
Collard, Gaspar,
Clergue, Escazu)

pMSSM: GAs as a tool for probing structure:

Interesting feature of GA's is the fitness distance correlation, and how it affects the behaviour of the population as it evolves. (Checked with MultiNest — Bayesian Inference — GA 10-100 x faster for CMSSM)

For this study use pMSSM, 23 parameters:

(Berger, Gainer, Hewett, Rizzo; Abdussalam, Allanach, Quevedo, Feroz, Hobson; Cahill-Rowley, Hewett, Ismail, Rizzo)

Observable	Value
$\left[\alpha_{\mathrm{EM}}(M_Z)^{\overline{MS}}\right]^{-1}$	127.950 ± 0.017
$\alpha_{\mathrm{S}}(M_Z)^{\overline{MS}}$	0.1185 ± 0.0006
$m_b(\mathrm{GeV})$	4.78 ± 0.06
$m_t(\mathrm{GeV})$	173.1 ± 0.6

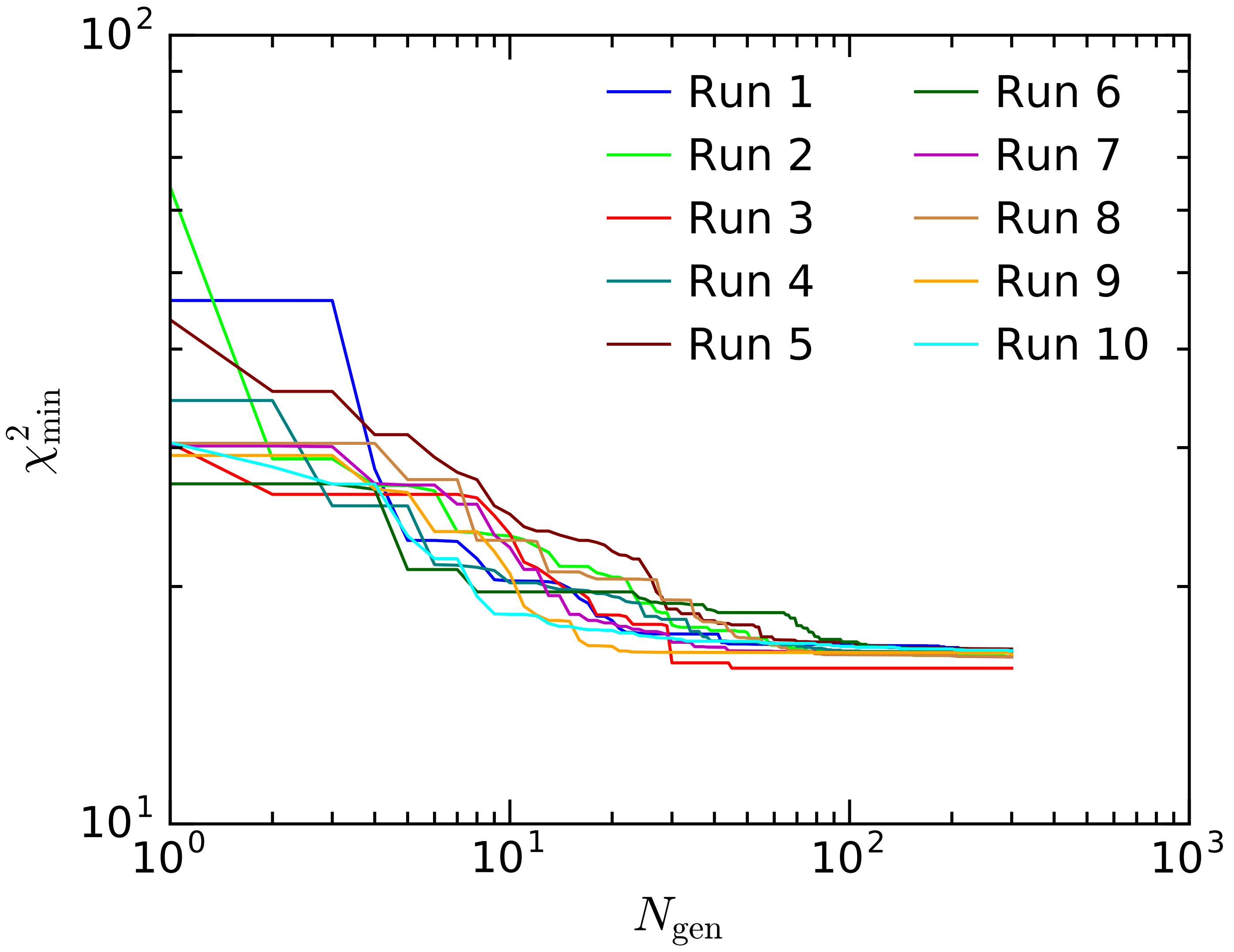
Parameter	Range
SM	
$\left[\alpha_{\mathrm{EM}}(M_Z)^{\overline{MS}}\right]^{-1}$	$[127.882, 128.018]$
$\alpha_{\mathrm{S}}(M_Z)^{\overline{MS}}$	$[0.1161, 0.1209]$
$m_b(\mathrm{GeV})$	$[4.54, 5.02]$
$m_t(\mathrm{GeV})$	$[170.1, 175.5]$
pMSSM (GUT scale)	
$M_1, M_2, M_3(\mathrm{GeV})$	[50,10000]
$m_{H_u}, m_{H_d}(\mathrm{GeV})$	[50,10000]
$m_{\tilde{Q}_{1,2}} m_{\tilde{Q}_3}(\mathrm{GeV})$	[50,10000]
$m_{\tilde{U}_{1,2}} m_{\tilde{U}_3}(\mathrm{GeV})$	[50,10000]
$m_{\tilde{D}_{1,2}} m_{\tilde{D}_3}(\mathrm{GeV})$	[50,10000]
$m_{\tilde{L}_{1,2}} m_{\tilde{L}_3}(\mathrm{GeV})$	[50,10000]
$m_{\tilde{E}_{1,2}} m_{\tilde{E}_3}(\mathrm{GeV})$	[50,10000]
$A_t, A_b, A_\tau(\mathrm{TeV})$	[-10,10]
$\tan\beta$	[2,62]

pMSSM: GAs as a tool for probing structure:

Fitness function is simply 1/likelihood derived from all experimental constraints: it singles out (g-2) of the muon as the offending observable.

$$\ln \mathcal{L}_{\rm Joint} \;=\; \ln \mathcal{L}_{\rm EWPO} + \ln \mathcal{L}_{B} + \ln \mathcal{L}_{\rm Higgs} + \ln \mathcal{L}_{\rm LEP} + \ln \mathcal{L}_{\rm LHC} + \ln \mathcal{L}_{\Omega_{\rm DM} h^2} + \ln \mathcal{L}_{\delta a_\mu^{\rm SUSY}}$$

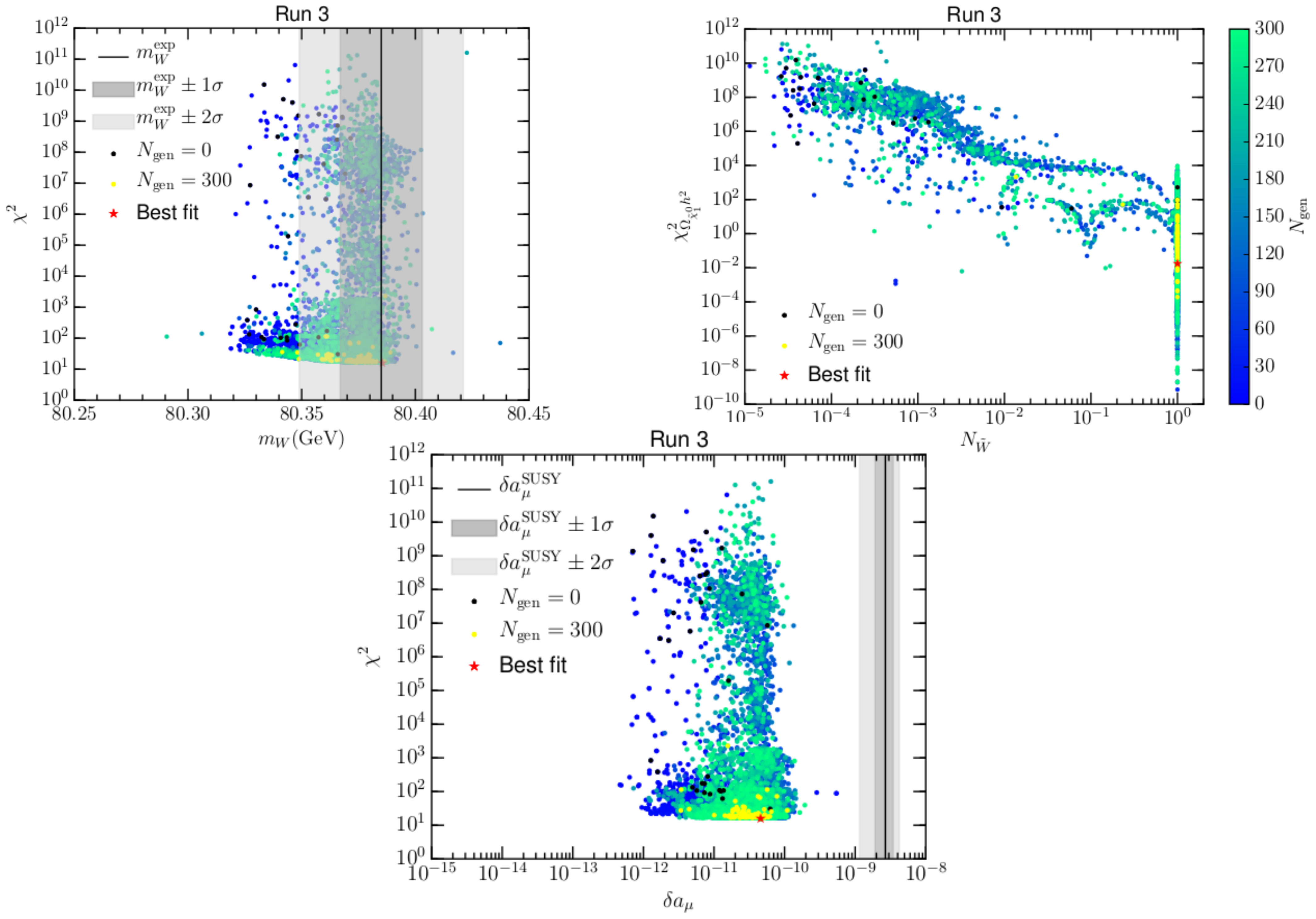
USED: **PIKAIA2.0** (Metcalf+Charbonneau), **SoftSUSY, FeynHiggs, ZFITTER, MicrOMEGAS, HiggSignals, PYTHIA, SModelS, NLL-Fast, Fastlim**.



	Run 1
$\chi^2_{\Omega_{\tilde\chi_1^0} h^2}$	0.0067
$\chi^2_{\rm HiggsSignals}$	1.2950
$\chi^2_{m_{h^0}}$	0.1125
$\chi^2_{M_W}$	0.1190
$\chi^2_{\sin^2\theta^{\rm lept}_{\rm eff}}$	0.1538
$\chi^2_{\Gamma_Z}$	0.0332
$\chi^2_{\Gamma_Z^{\rm inv}}$	2.3054
$\chi^2_{BR(B\to X_s\gamma)}$	0.0664
$\chi^2_{BR(B_s^0\to\mu^+\mu^-)}$	0.1647
$\chi^2_{\frac{BR(B_u\to\tau\nu)}{BR(B_u\to\tau\nu)_{\rm SM}}}$	0.0140
$\chi^2_{\rm LEP}$	0.0000
$\chi^2_{\rm LHC}$	0.0000
$\chi^2_{\delta a_\mu^{\rm SUSY}}$	12.2691
$\chi^2_{\rm tot}$	16.5398

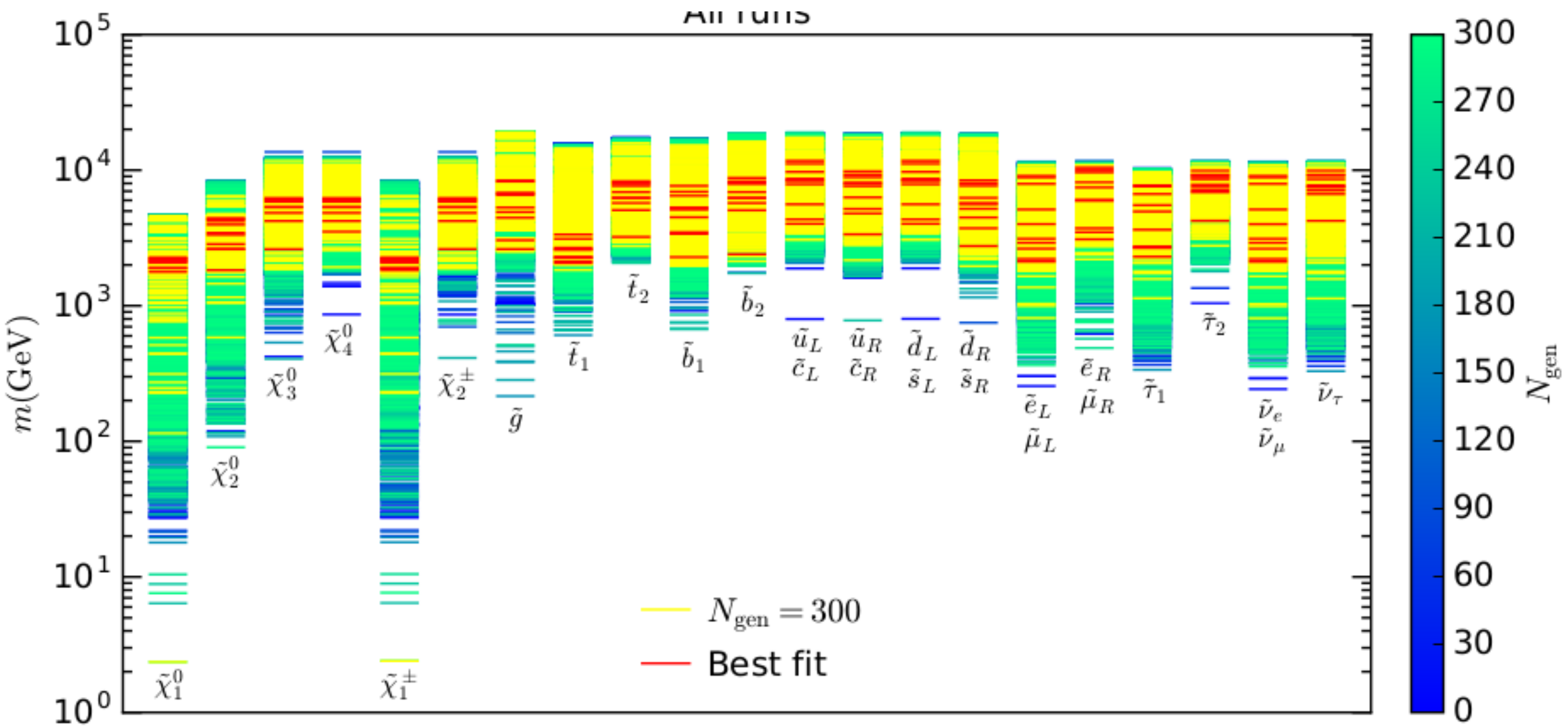
pMSSM: GAs as a tool for probing structure:

Information about the structure can be inferred from the “flow” (assuming fitness distance correlation). e.g. the W mass is easy to fit and not constraining, DM is hard and constraining, g-2 is impossible.



pMSSM: GAs as a tool for probing structure:

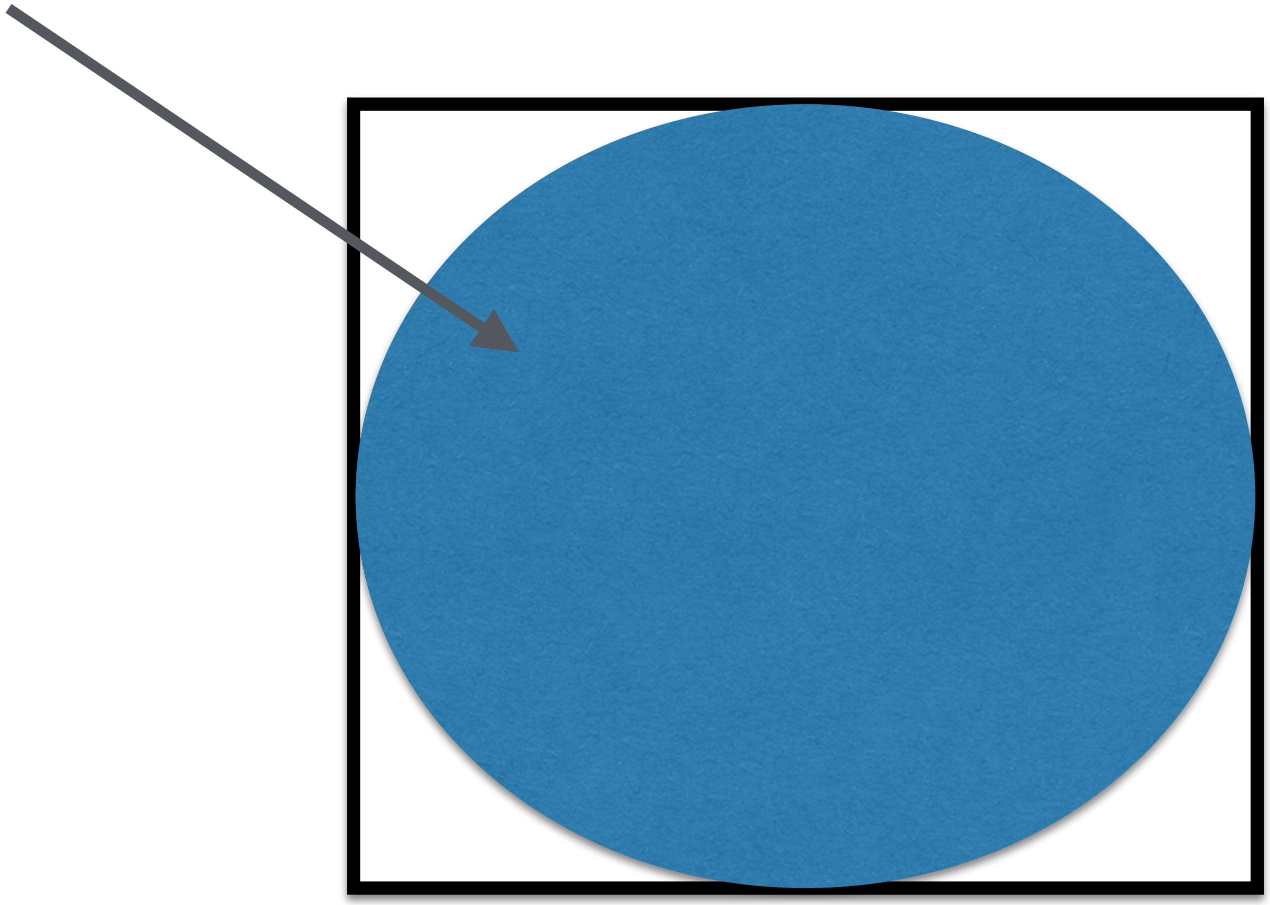
You can get “predictions” from the final generations. e.g. in this case the spectrum:



pMSSM: GAs as a tool for probing structure:

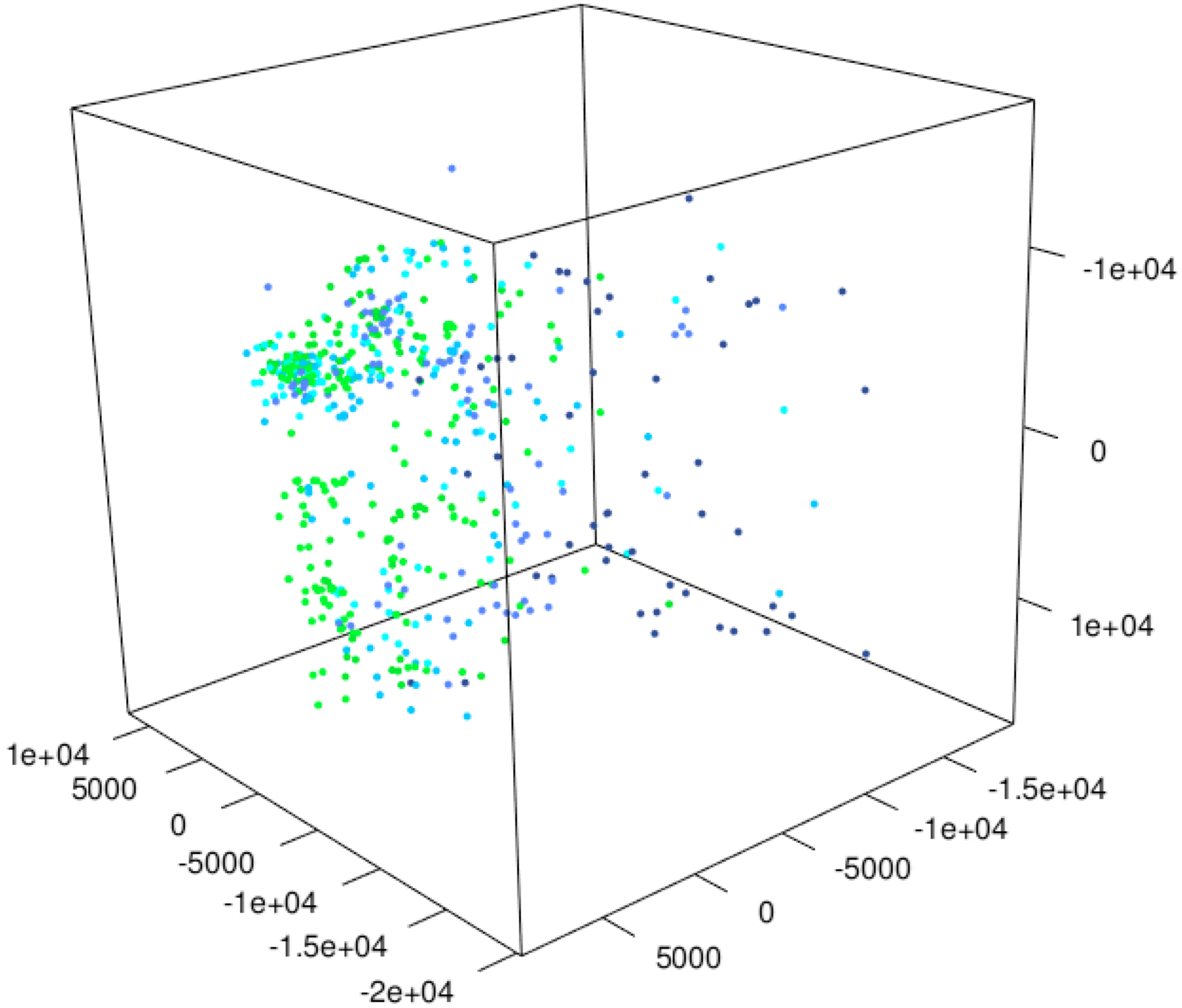
Note the “large dimensionality problem”: in 19 dimensions, slices give a misleading representation of the structure

In 19D this ball occupies only 10^(-7) of the volume of the cube!



pMSSM: GAs as a tool for probing structure:

Slices give a good idea of the *flow*, but non-linear (Sammon) mapping gives a better image of the *clustering*:



Conclusions

- GA’s are a promising method of searching for favourable string vacua
- Search difficulty appears to increase logarithmically with difficulty => 10^500 is doable!!
- Fitness distance correlation important (The problem cannot be a needle in a haystack)
- pMSSM studies suggest interesting approach to study string landscape structure
- But need to decide what you want to ask