Generalization of Hard and Soft Constraints

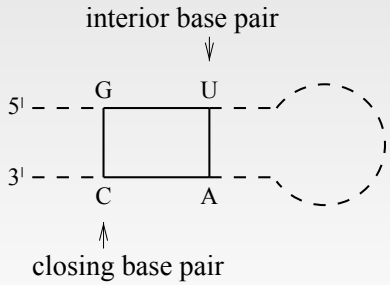
Ronny Lorenz
ronny@tbi.univie.ac.at

Institute for Theoretical Chemistry
University of Vienna

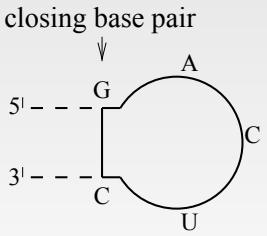
Bled, Slovenia, February 14, 2013

RNA Secondary structure prediction

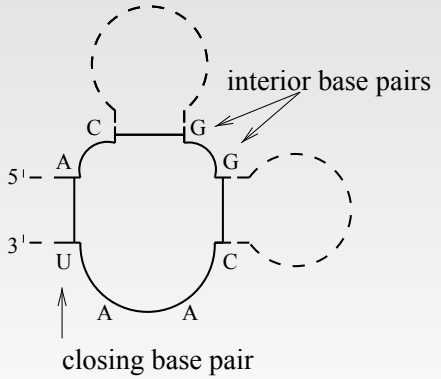
Secondary structures can be uniquely decomposed into loops



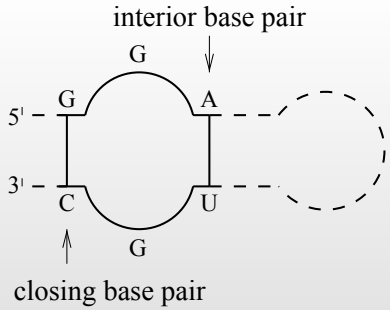
stacking pair



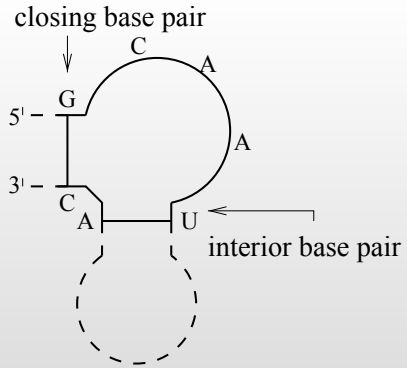
hairpin loop



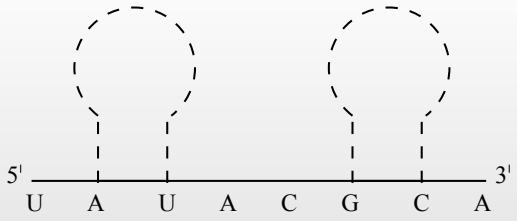
multi loop



interior loop

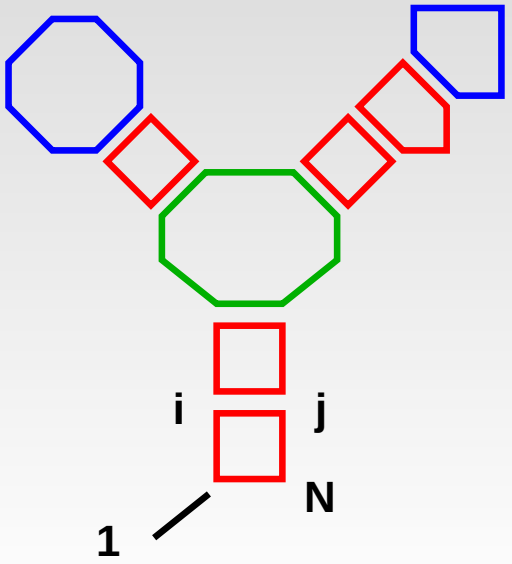


bulge



exterior loop

RNA Secondary structure prediction

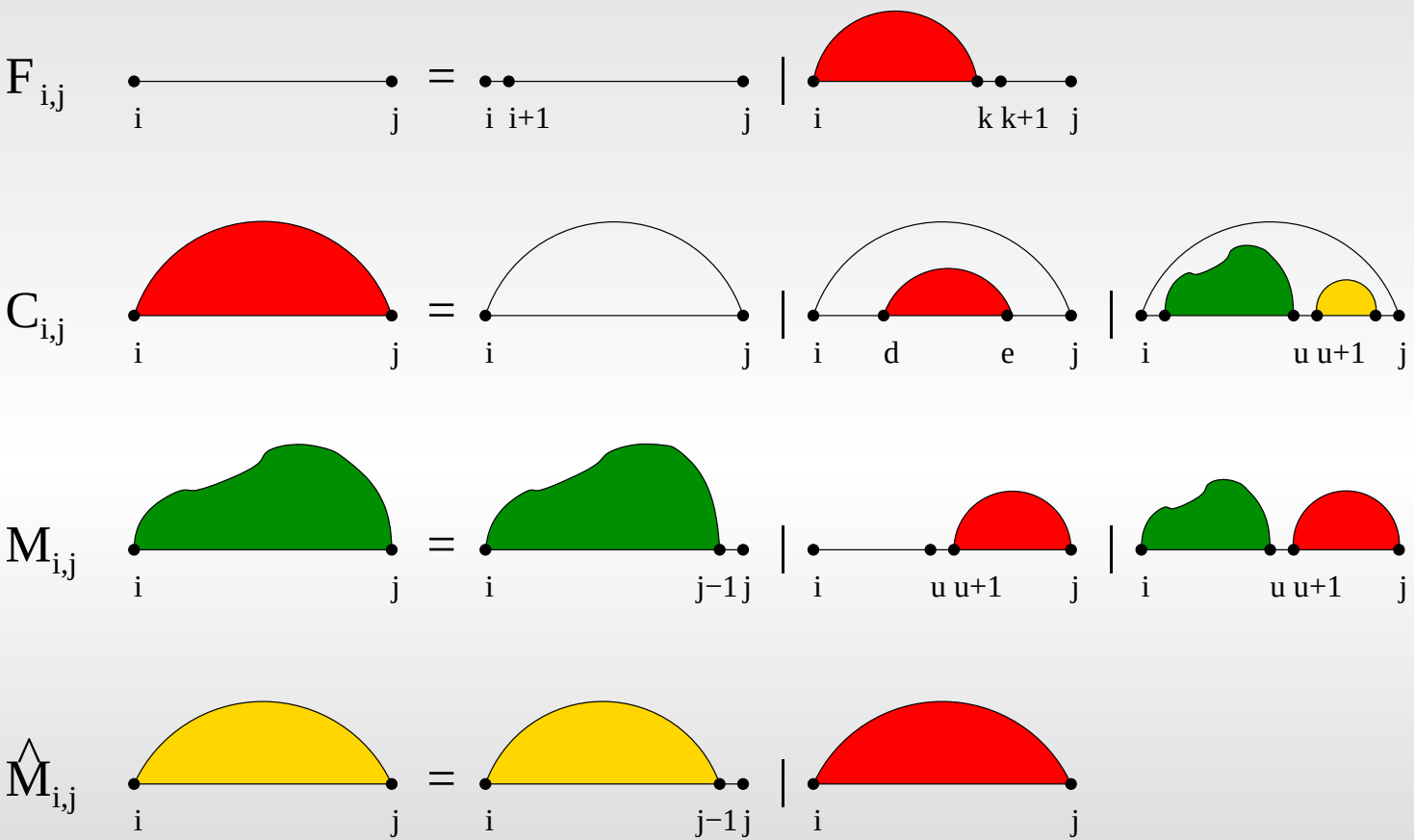


$$E(\mathcal{S}) = \sum_{L \in \mathcal{S}} E(L)$$

- The free energy of a secondary structure is the sum of the free energy of the loops its composed of
- Loop energies depend on loop type, loop size and sequence
- Energy parameters are measured experimentally or extrapolated by mathematical models

RNA Secondary structure prediction

Decomposition scheme



What is constraint folding

What happens during secondary structure prediction:

- decomposition scheme is applied to a sequence
- underlying energy model assign contributions to each decomposition
- algorithm finds e.g. an optimal structure (MFE) or adds up Boltzmann factors (PF)

What is constraint folding

What happens during secondary structure prediction:

- decomposition scheme is applied to a sequence
- underlying energy model assign contributions to each decomposition
- algorithm finds e.g. an optimal structure (MFE) or adds up Boltzmann factors (PF)

But:

- the energy model is not perfect
- experiment (e.g. SHAPE) may suggest differently to a prediction
- environmental settings may exclude a set of structures

What is constraint folding

What happens during secondary structure prediction:

- decomposition scheme is applied to a sequence
- underlying energy model assign contributions to each decomposition
- algorithm finds e.g. an optimal structure (MFE) or adds up Boltzmann factors (PF)

But:

- the energy model is not perfect
- experiment (e.g. SHAPE) may suggest differently to a prediction
- environmental settings may exclude a set of structures

Secondary structure constraints:

- disallow certain parses of the decomposition scheme (Hard Constraints)
  Example: exclude a (set of) nucleotide(s) from base pairing
- alter the energy contributions of the model (Soft Constraints)
  Example: add a bonus/malus when a nucleotide is considered unpaired

What is constraint folding

Hard Constraints allows for cutting out/ inserting[1] points in the secondary structure energy landscape

[1] circumvention of build-in constraints, e.g canonical base pairs

What is constraint folding

Hard Constraints allows for cutting out/ inserting[1] points in the secondary structure energy landscape

[2]

[1] circumvention of build-in constraints, e.g canonical base pairs

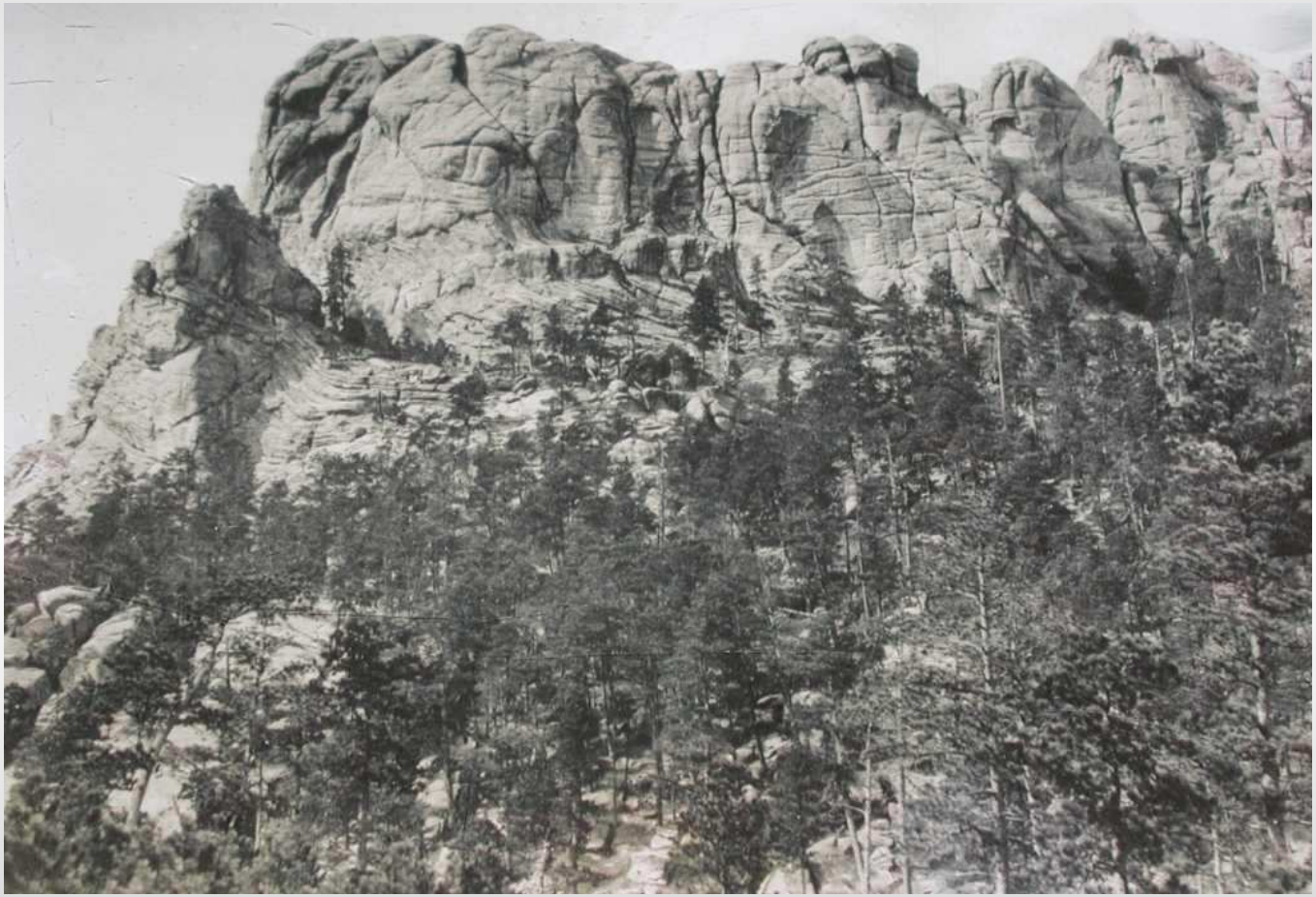
[2] Gobierno de Álvaro Colom, Guatemala

What is constraint folding

Soft Constraints allow for shifting points in the landscape up or down

What is constraint folding

Soft Constraints allow for shifting points in the landscape up or down



Mount Rushmore 1925

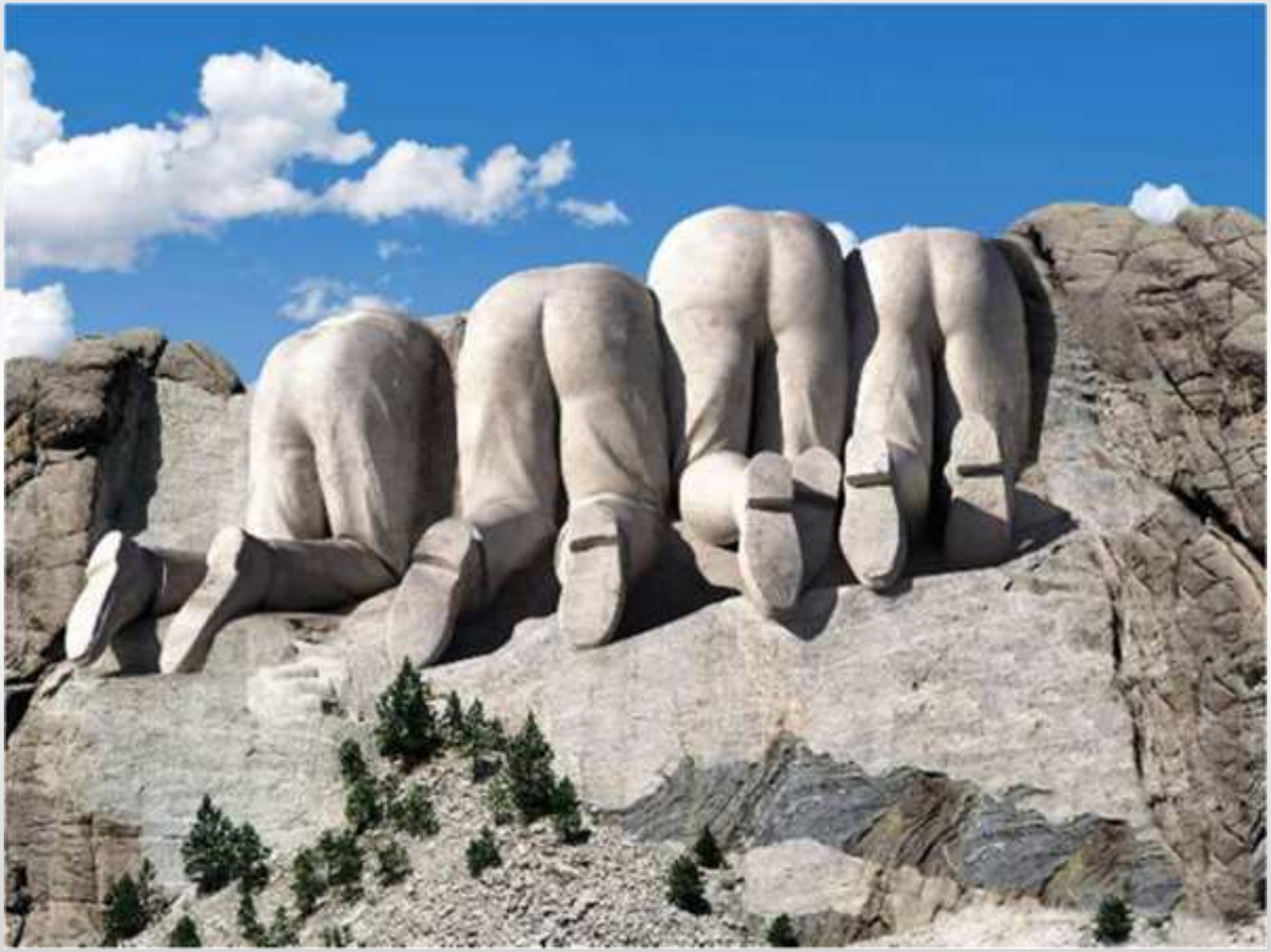
What is constraint folding

Soft Constraints allow for shifting points in the landscape up or down

Mount Rushmore Today

What is constraint folding

Soft Constraints allow for shifting points in the landscape up or down



Mount Rushmore from the back

What is constraint folding

Secondary structure constraints aware programs:

- UNAfold[3] (hard)
- RNAstructure[4] (hard, soft)
- RNApbfold[5] (hard, soft)
- ViennaRNA Package[6] (hard)

[3] Markham et al., 2008
[4] Reuter et al., 2010
[5] Washietl S. et al., 2012
[6] Hofacker et al., 1994, Lorenz et al. 2011

What is constraint folding

Secondary structure constraints aware programs:

- UNAfold[3] (hard)
- RNAstructure[4] (hard, soft)
- RNApbfold[5] (hard, soft)
- ViennaRNA Package[6] (hard)

Are the above implementations sufficient?

---

[3] Markham et al., 2008
[4] Reuter et al., 2010
[5] Washietl S. et al., 2012
[6] Hofacker et al., 1994, Lorenz et al. 2011

What is constraint folding

Secondary structure constraints aware programs:

- UNAfold[3] (hard)
- RNAstructure[4] (hard, soft)
- RNApbfold[5] (hard, soft)
- ViennaRNA Package[6] (hard)

Are the above implementations sufficient?

Of course NOT, so we need some generalization!

---

[3] Markham et al., 2008
[4] Reuter et al., 2010
[5] Washietl S. et al., 2012
[6] Hofacker et al., 1994, Lorenz et al. 2011

Generalized Hard constraints

Discriminate between the decomposition steps (loop types)

1. Do something about base pairs:
   - (dis)allow particular base pair to appear in exterior-, hairpin-, interior-, multibranch-loops
   - distinguish between enclosing and enclosed base pairs
2. Do something with unpaired nucleotides:
   - specify whether or not a nucleotide may be unpaired in a distinguished loop type

Example: Base pair (i,j) has to pair but may only enclose a multiloop

Hard constraints can be expressed in terms of a boolean function

$$f(\vec{x}, d, data) = 0|1$$

with nucleotide position vector $\vec{x}$, decomposition step d and some *data* compound that decides whether or not a decomposition step is processed.

Generalized Soft constraints

Generalization similar to Hard constraints

1. Discriminate between the decomposition steps
2. Generalize to a pseudo-energy function

$$f(\vec{x}, d, data) = e$$

to obtain bonus/malus for a particular decomposition step

Generalized Soft constraints

Generalization similar to Hard constraints

1. Discriminate between the decomposition steps
2. Generalize to a pseudo-energy function

$$f(\vec{x}, d, data) = e$$

to obtain bonus/malus for a particular decomposition step

What do we gain with those generalizations?

An example application for generalized soft constraints

RNA2Dfold

- classified dynamic programming approach
- computes MFE or partition function for a set of distance classes
- a distance class is the set of all structures whith a specified base pair distance to two initially chosen reference structures
  E.g. with reference structures s_1 and s_2, the distance class $(5, 17)$ is populated with all structures s that fulfill

$$d_{BP}(s, s_1) = 5 \wedge d_{BP}(s, s_2) = 17$$

- underlying algorithm is rather slow in terms of asymptotic time complexity $O(N^7)$ and consumes a lot of memory $O(N^4)$

An example application for generalized soft constraints

RNA2Dfold

- classified dynamic programming approach
- computes MFE or partition function for a set of distance classes
- a distance class is the set of all structures whith a specified base pair distance to two initially chosen reference structures
  E.g. with reference structures s_1 and s_2, the distance class $(5, 17)$ is populated with all structures s that fulfill

$$d_{BP}(s, s_1) = 5 \wedge d_{BP}(s, s_2) = 17$$

- underlying algorithm is rather slow in terms of asymptotic time complexity $O(N^7)$ and consumes a lot of memory $O(N^4)$
- However, it can be used for e.g. barrier heuristics, metastable state detection or even for RNA folding kinetics computations (see my talk last year)

An example application for generalized soft constraints

RNA2Dfold

- classified dynamic programming approach
- computes MFE or partition function for a set of distance classes
- a distance class is the set of all structures whith a specified base pair distance to two initially chosen reference structures
  E.g. with reference structures s_1 and s_2, the distance class $(5, 17)$ is populated with all structures s that fulfill

$$d_{BP}(s, s_1) = 5 \wedge d_{BP}(s, s_2) = 17$$

- underlying algorithm is rather slow in terms of asymptotic time complexity $O(N^7)$ and consumes a lot of memory $O(N^4)$
- However, it can be used for e.g. barrier heuristics, metastable state detection or even for RNA folding kinetics computations (see my talk last year)

Question: Can this be done more efficiently?

Distortion of the energy landscape

Idea: Approximation of the RNA2Dfold distance classes

- Sample structures from the whole Boltzmann ensemble $O(n^3)$
- classify each sample according to two chosen reference structures
- retrieve the MFE representative of the sampled distance classes
- compute partition function for each resulting distance classes

Drawback:

Sampling would only retrieve structure states from the lower portion of the energy landscape

Distortion of the energy landscape

Use generalized Soft constraints to favorize structures according to their distance to the chosen reference structures s_1 and s_2

$$
\begin{aligned}
Q &= \sum_s \exp^{-E(s)/RT} \\
Q^{distorted} &= \sum_s f(\mathcal{S}, s, s_1, s_2) \cdot \exp^{-E(s)/RT} \\
&= \sum_s x^{d_{BP}(s,s_1)} \cdot y^{d_{BP}(s,s_2)} \cdot \exp^{-E(s)/RT}
\end{aligned}
$$

In pseudo energy notations with $x = \exp^{-x'/RT}$ and $y = \exp^{-y'/RT}$

$$
Q^{distorted} = \sum_s \exp^{-(E(s)+x' \cdot d_{BP}(s,s_1)+y' \cdot d_{BP}(s,s_2))/RT}
$$

Distortion the energy landscape

Now, choose x and y such that s_1 and s_2 and the MFE structure s_{MFE} are equally probable.

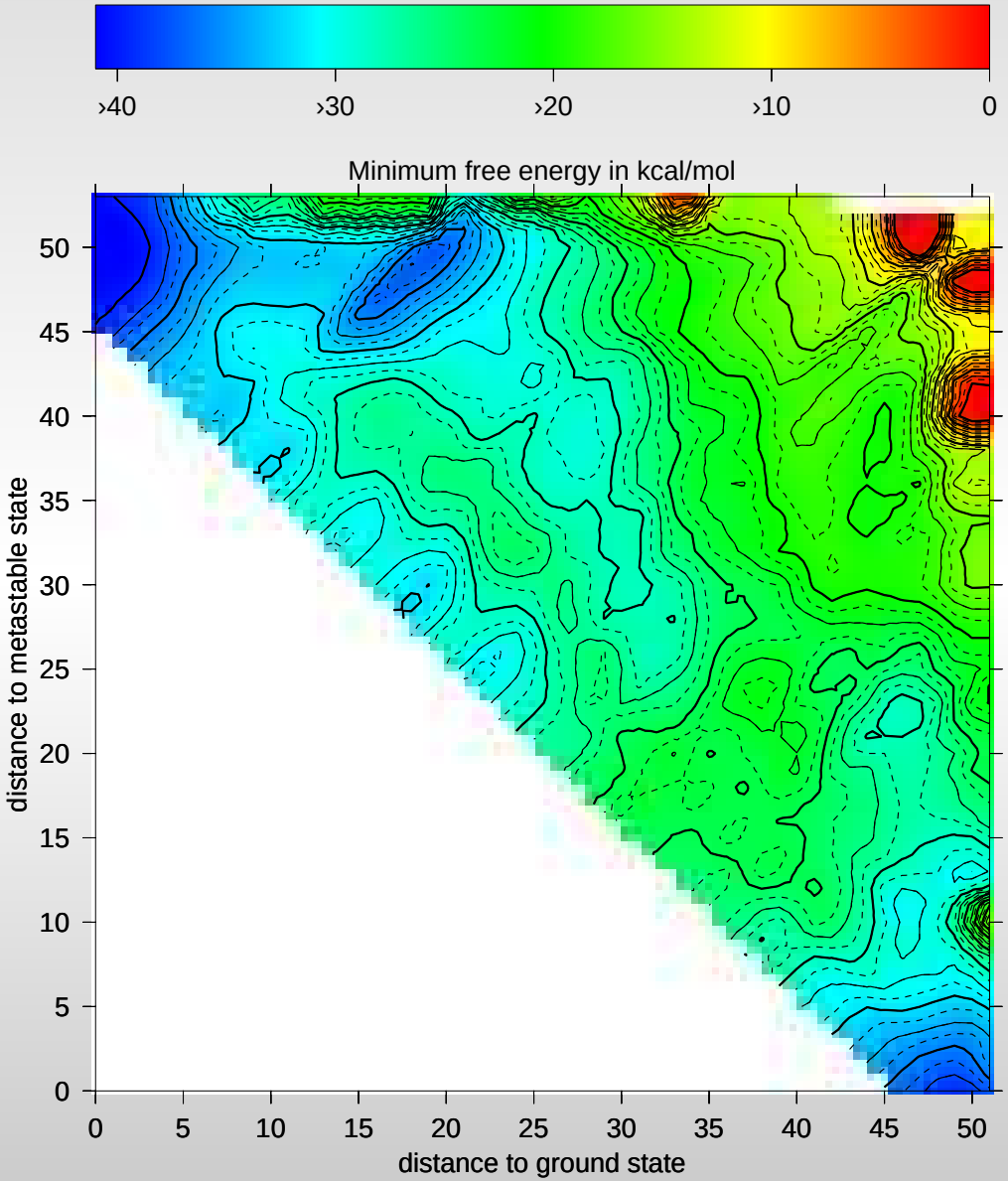
$$P(s) = \frac{\exp^{-E(s)/RT}}{Q}$$

$$\begin{aligned}
\exp^{-E(s_1)/RT} \cdot x^0 \cdot y^{d_{BP}(s_1,s_2)} &= \exp^{-E(s_2)/RT} \cdot x^{d_{BP}(s_1,s_2)} \cdot y^0 \\
\exp^{-E(s_1)/RT} \cdot x^0 \cdot y^{d_{BP}(s_1,s_2)} &= \exp^{-E(s_{MFE})/RT} \cdot x^{d_{BP}(s_1,s_{MFE})} \cdot y^{d_{BP}(s_2,s_{MFE})} \\
\exp^{-E(s_2)/RT} \cdot x^{d_{BP}(s_1,s_2)} \cdot y^0 &= \exp^{-E(s_{MFE})/RT} \cdot x^{d_{BP}(s_1,s_{MFE})} \cdot y^{d_{BP}(s_2,s_{MFE})}
\end{aligned}$$

The above equations can now be solved for x and y

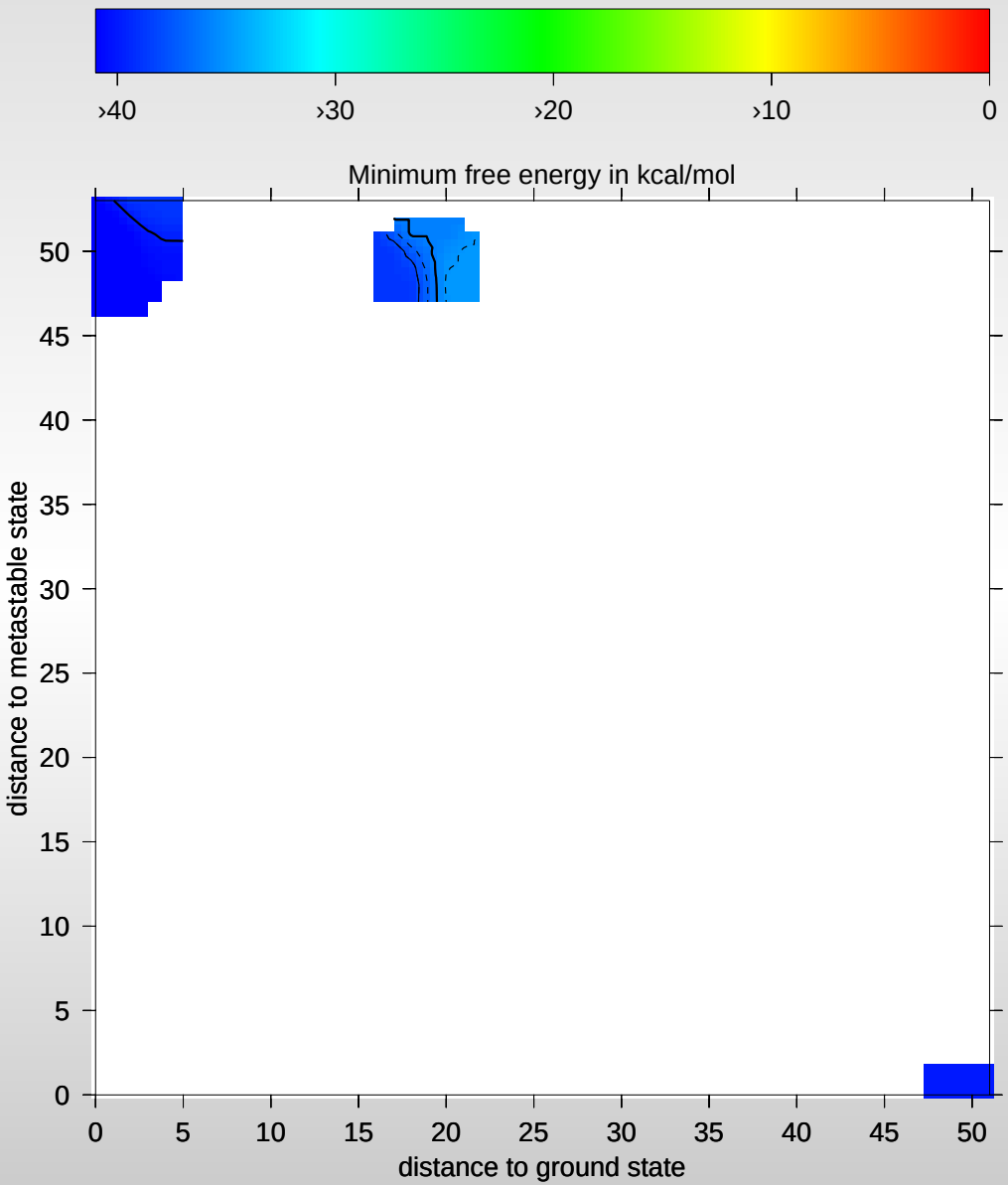
And then sample from this distorted landscape

Example 1: designed RNA switch [7] - RNA2Dfold

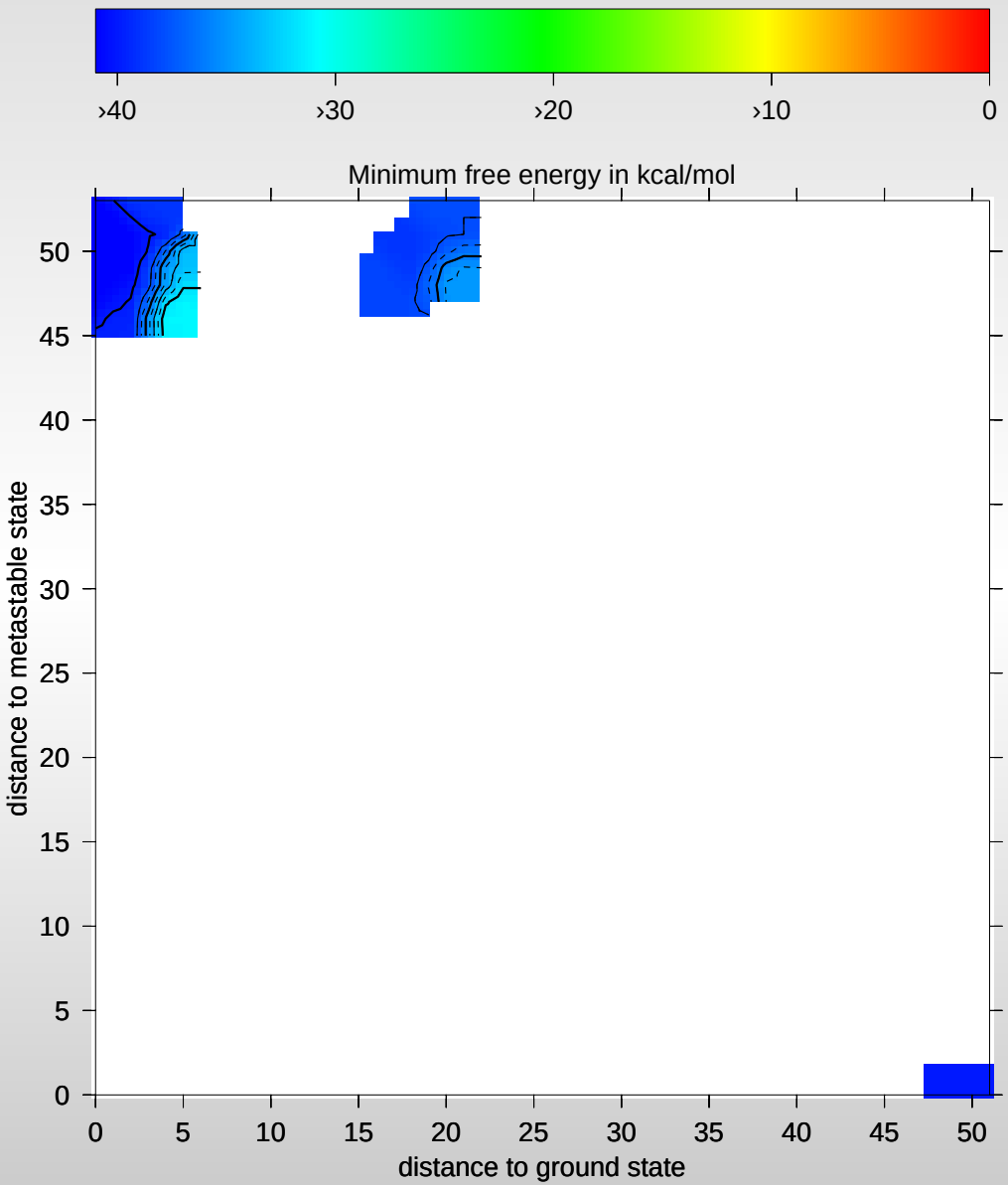


[7] Xayaphoummine et al. 2007

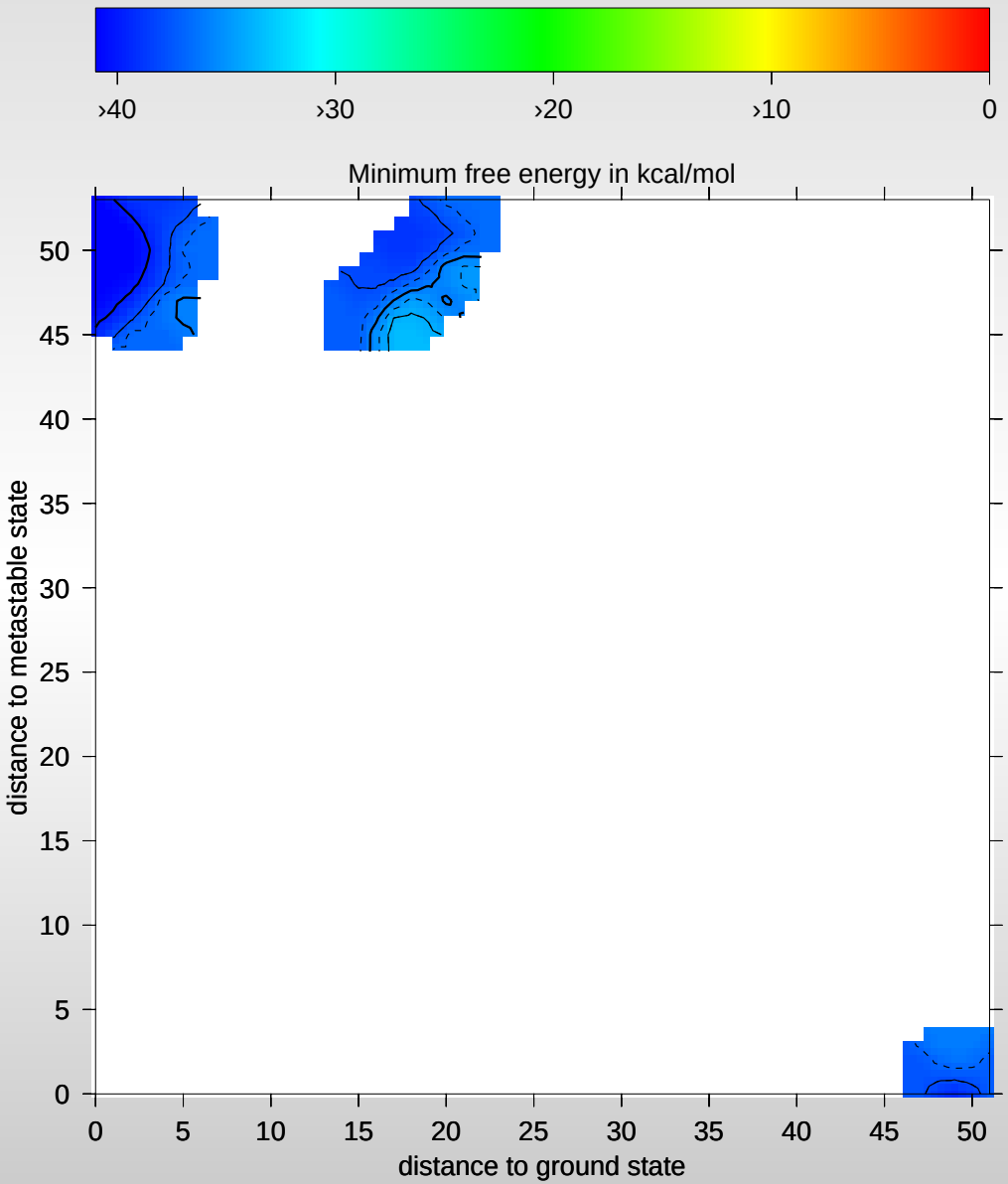
Example 1: designed RNA switch - RNAsubopt -p 10^2



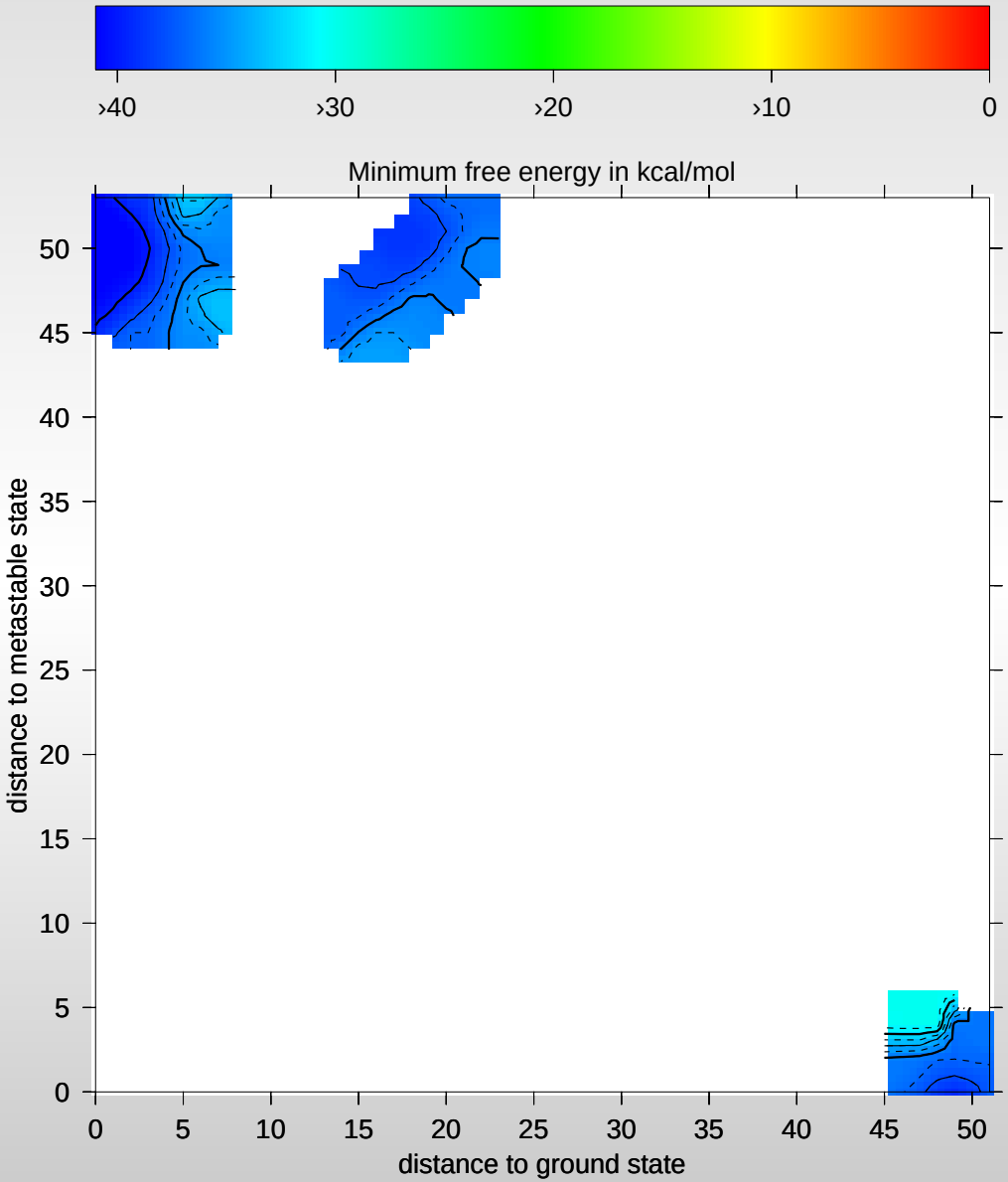
Example 1: designed RNA switch - RNAsubopt -p 10^3



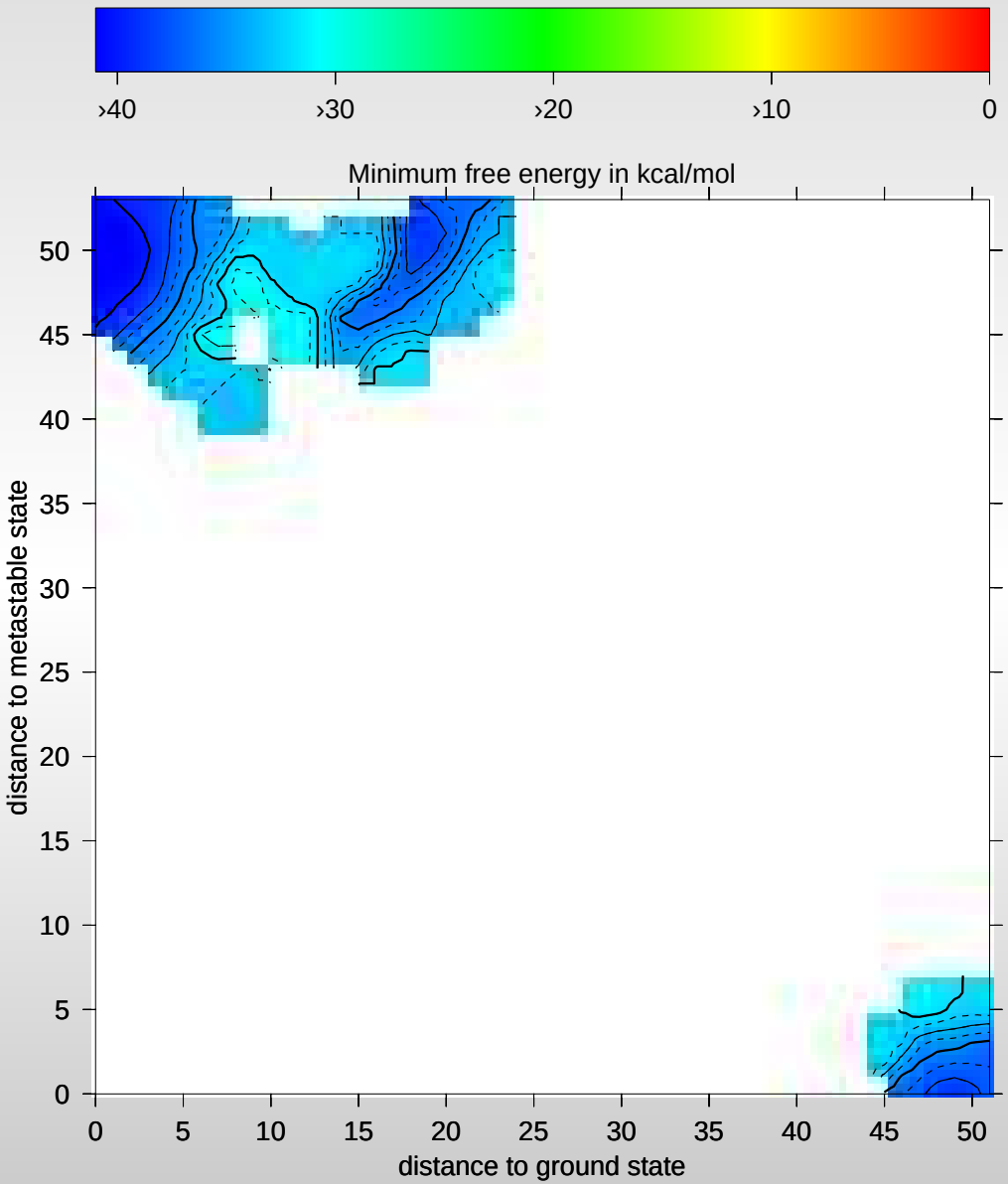
Example 1: designed RNA switch - RNAsubopt -p 10^4



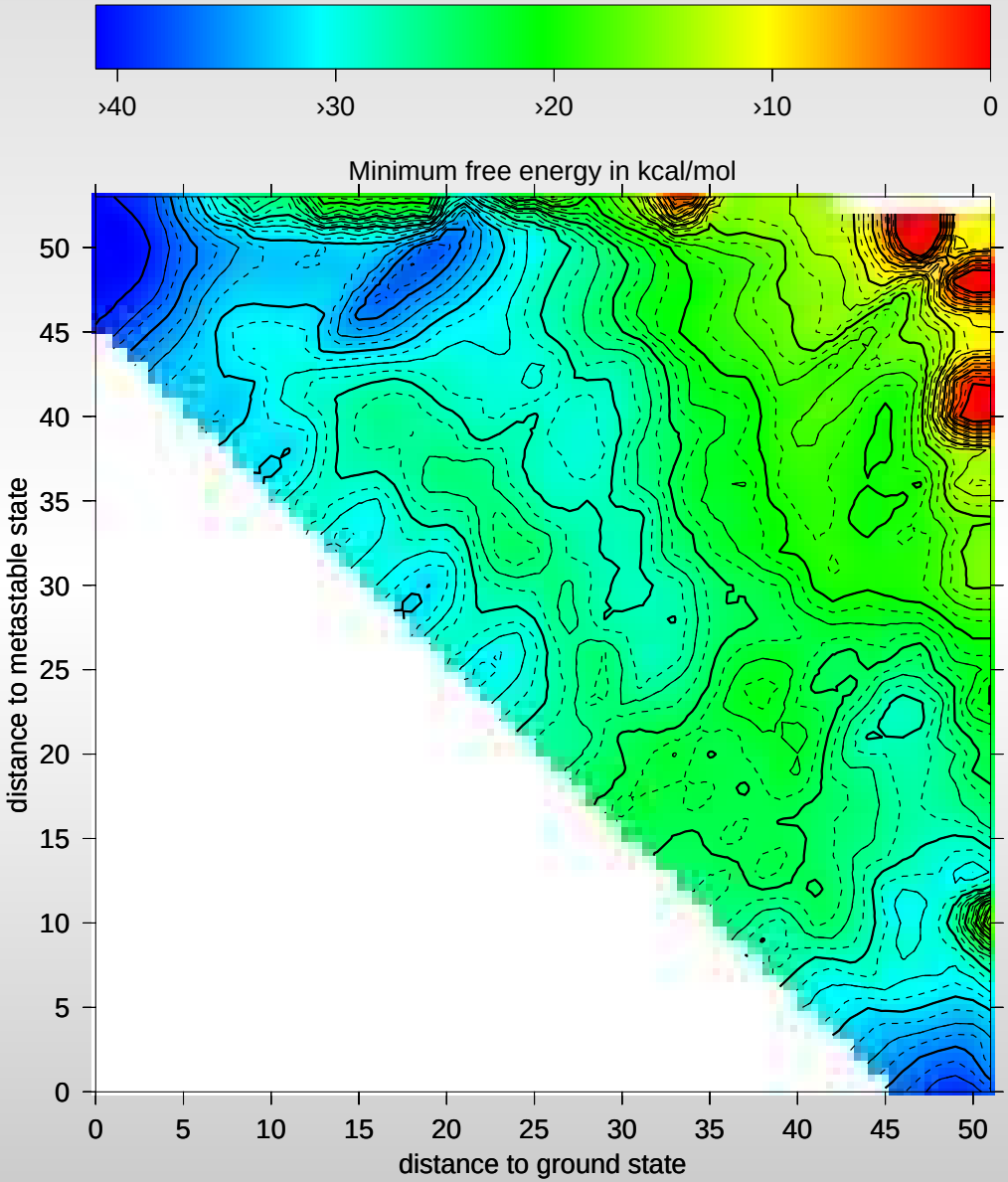
Example 1: designed RNA switch - RNAsubopt -p 10^5



Example 1: designed RNA switch - RNAsubopt -p 10^6

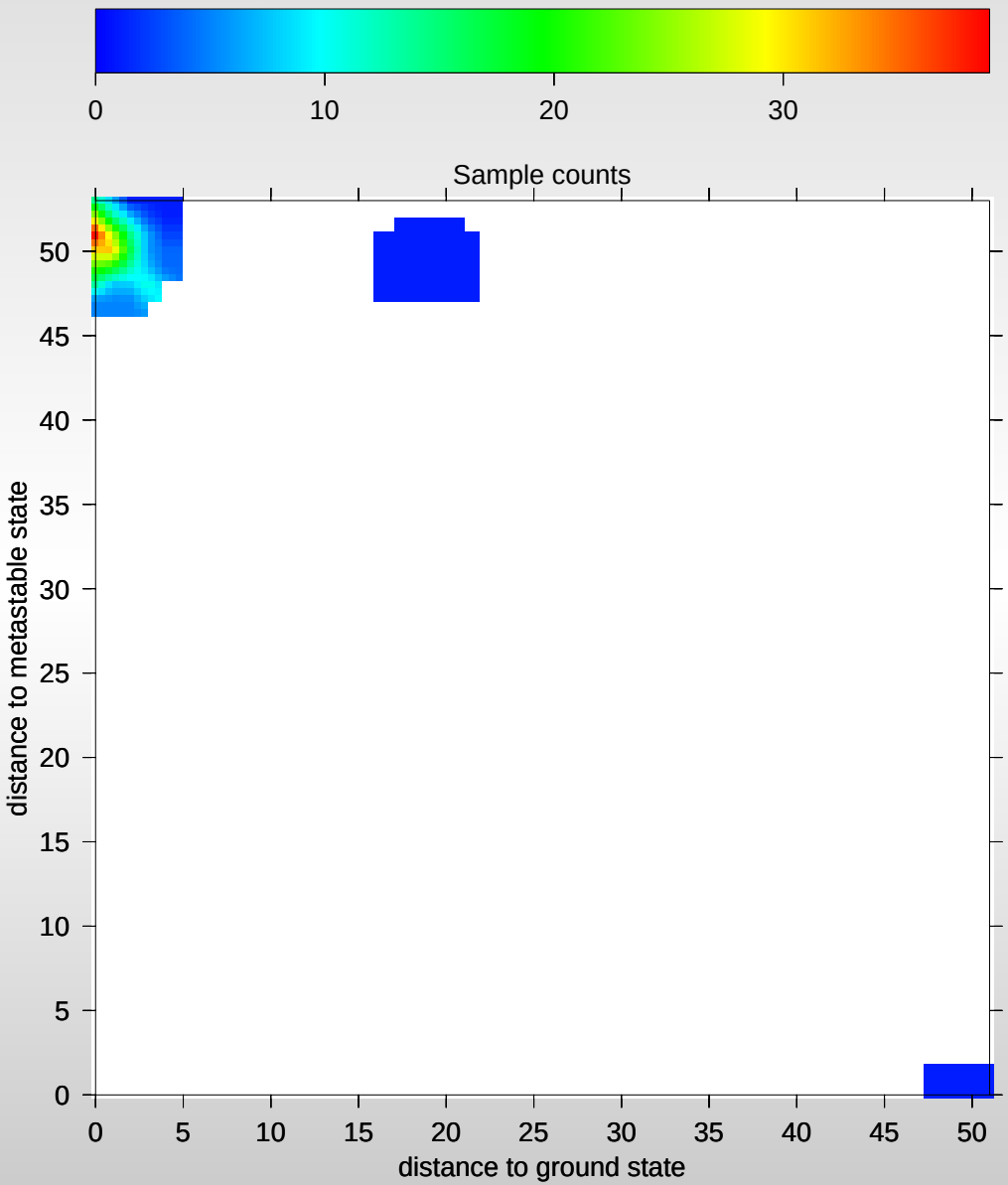


Example 1: designed RNA switch [8] - RNA2Dfold

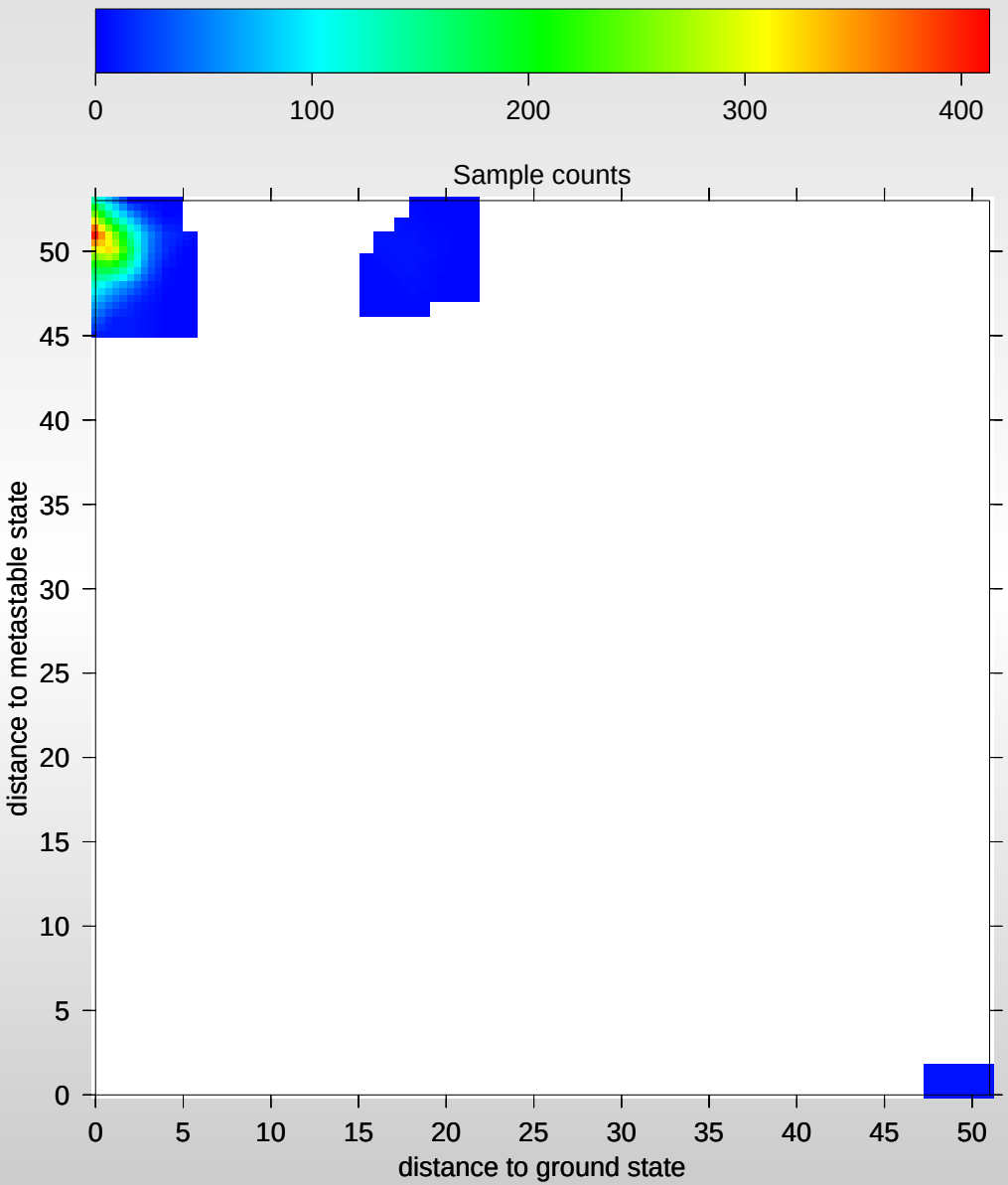


[8] Xayaphoummine et al. 2007

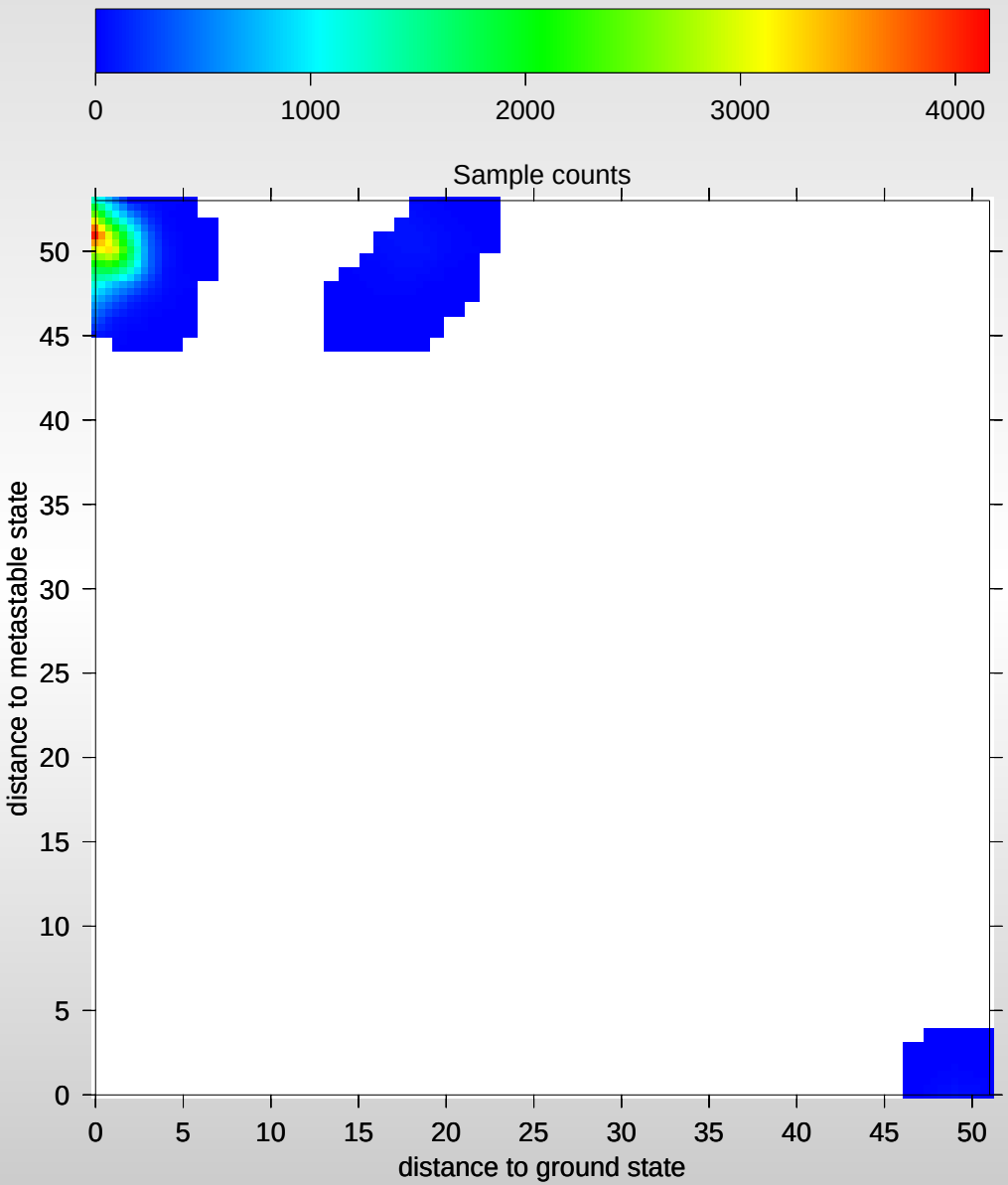
Example 1: designed RNA switch - RNAsubopt -p 10^2



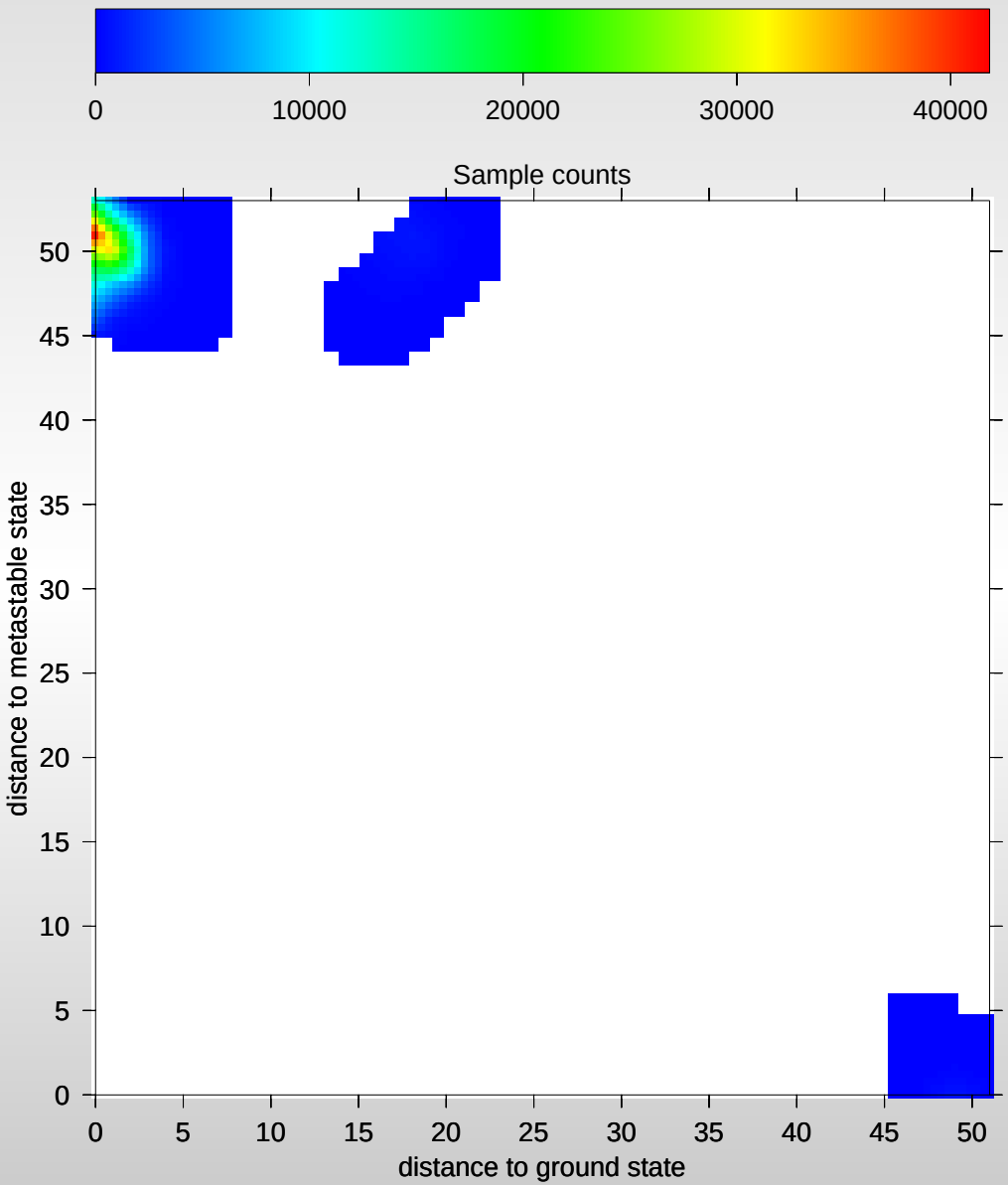
Example 1: designed RNA switch - RNAsubopt -p 10^3



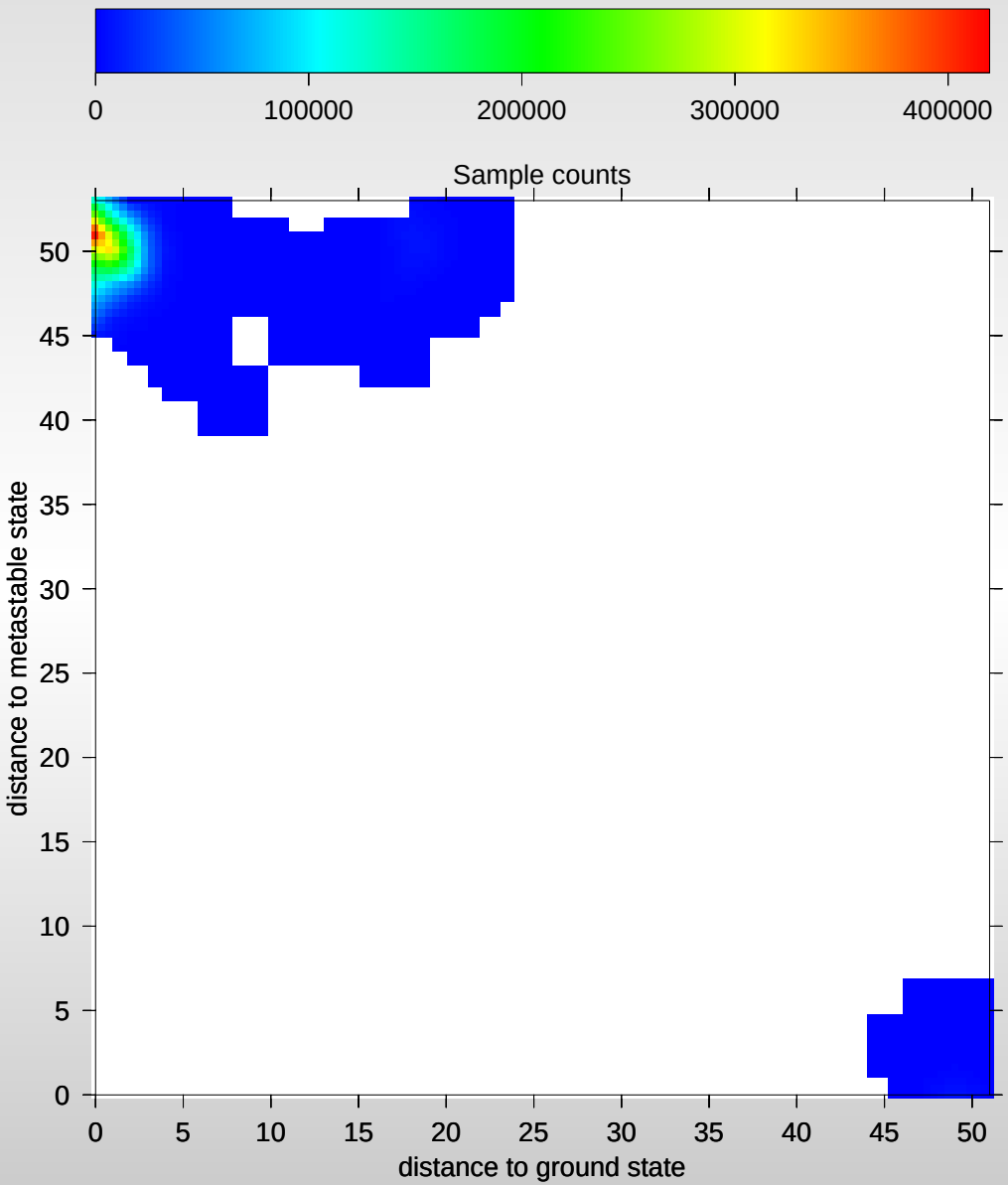
Example 1: designed RNA switch - RNAsubopt -p 10^4



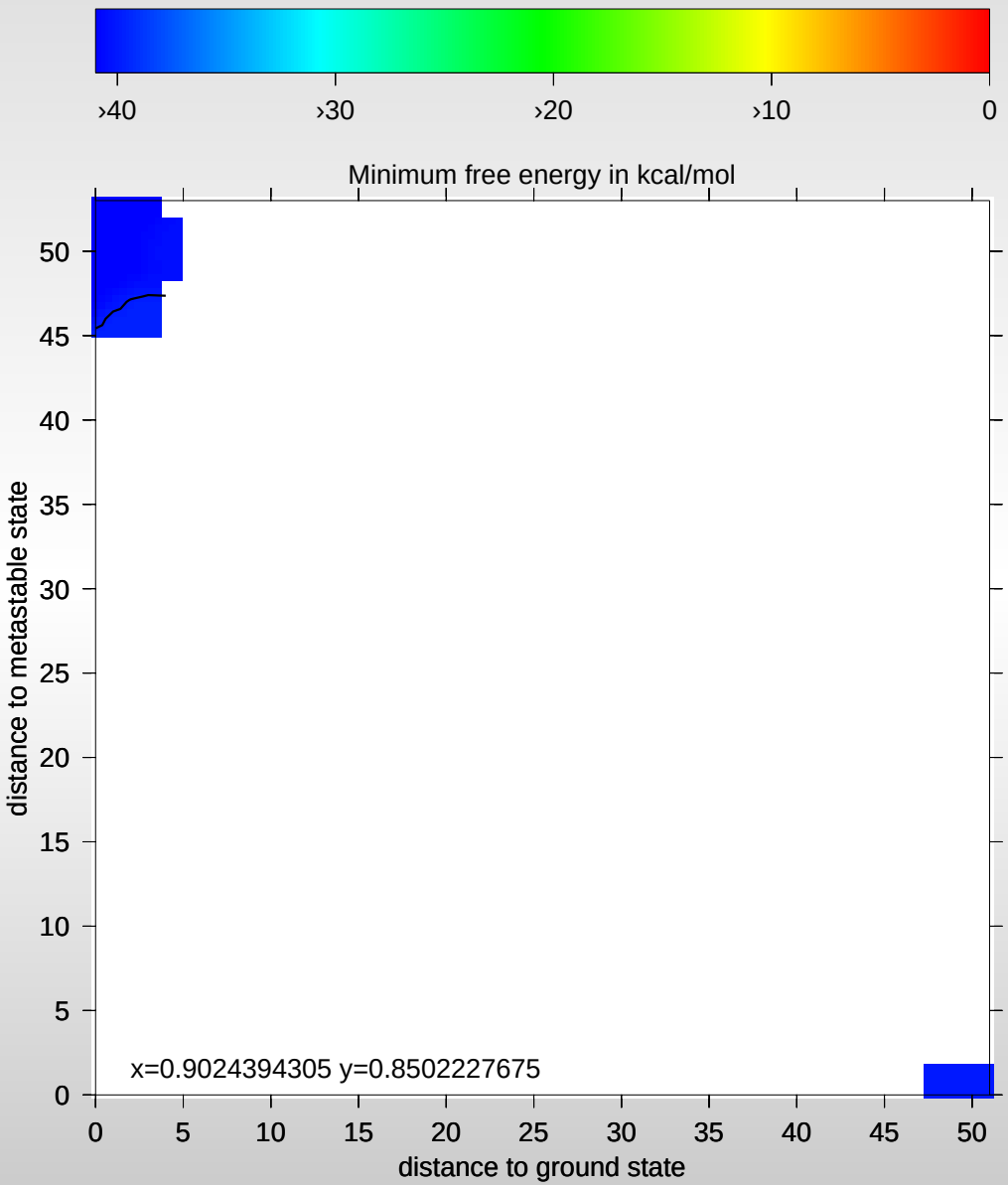
Example 1: designed RNA switch - RNAsubopt -p 10^5



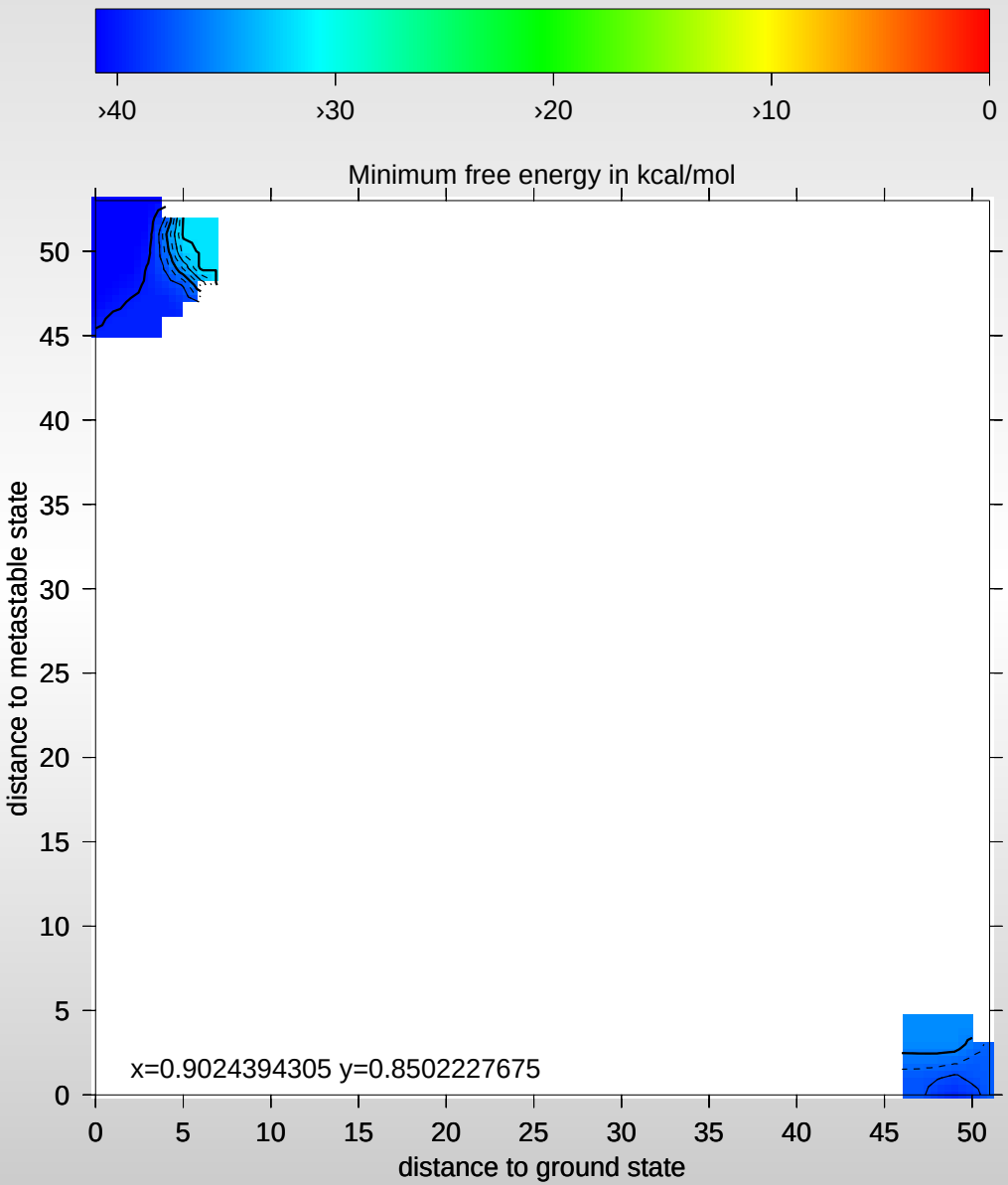
Example 1: designed RNA switch - RNAsubopt -p 10^6



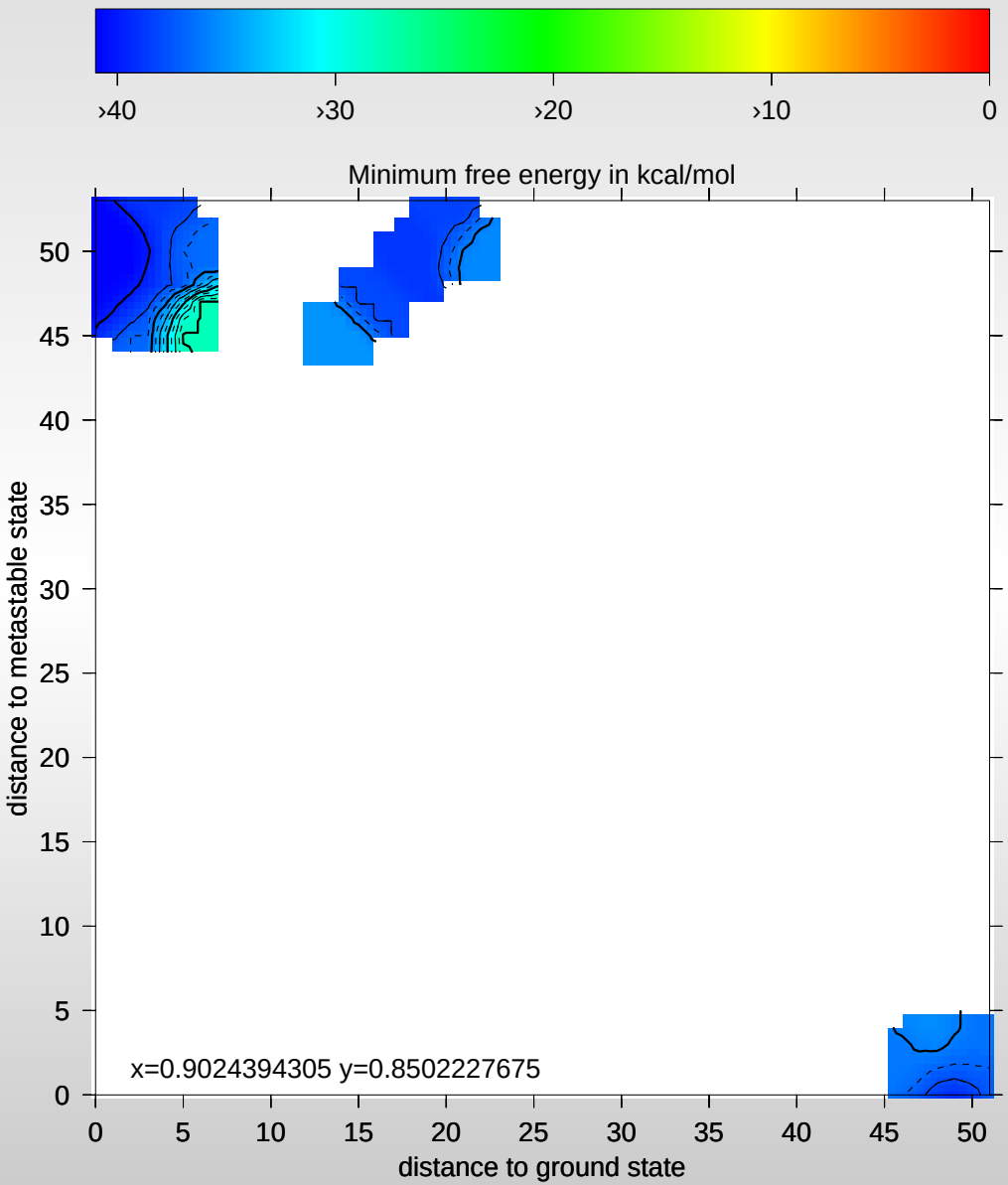
Example 1: designed RNA switch - distortion 10^2



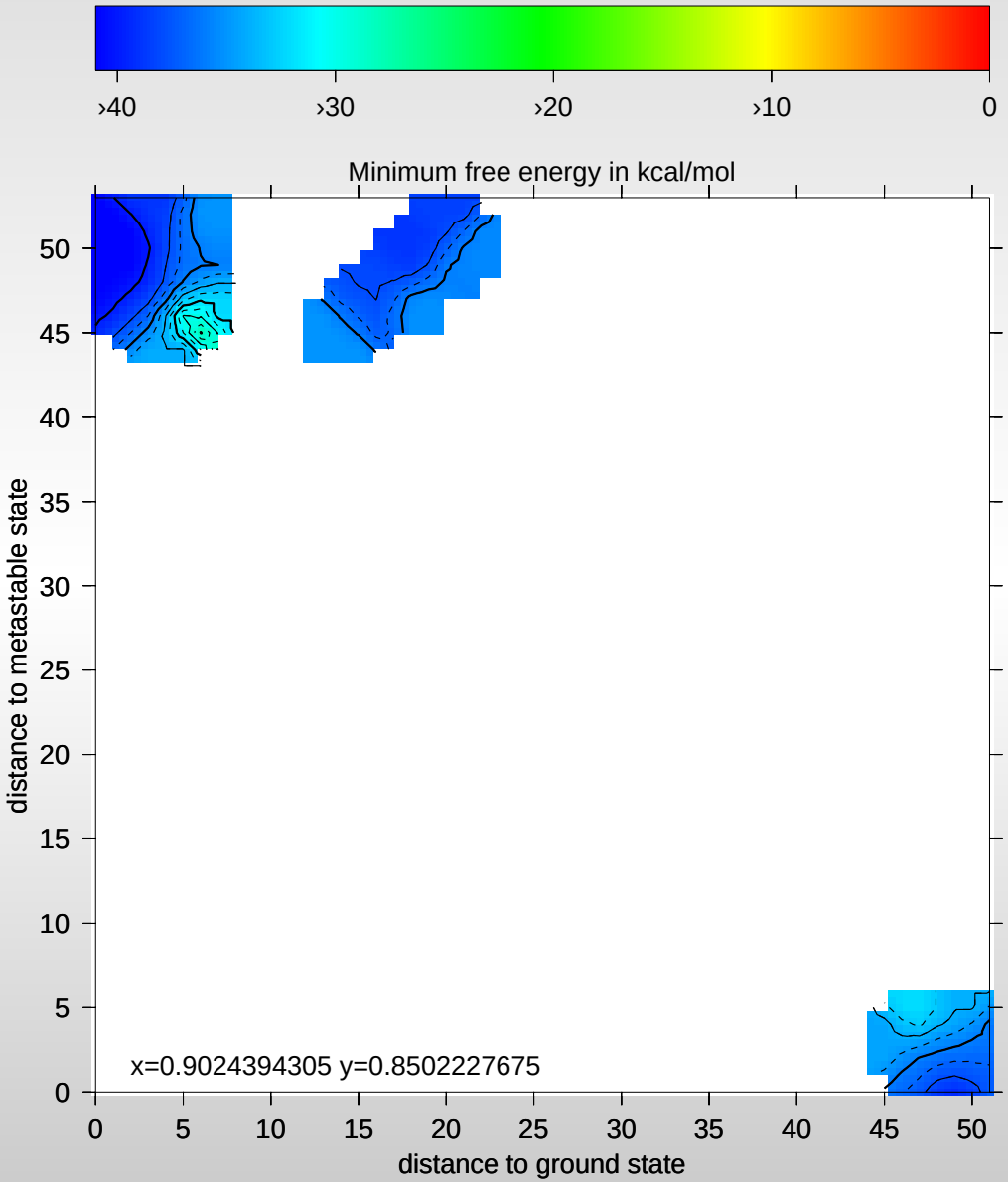
Example 1: designed RNA switch - distortion 10^3



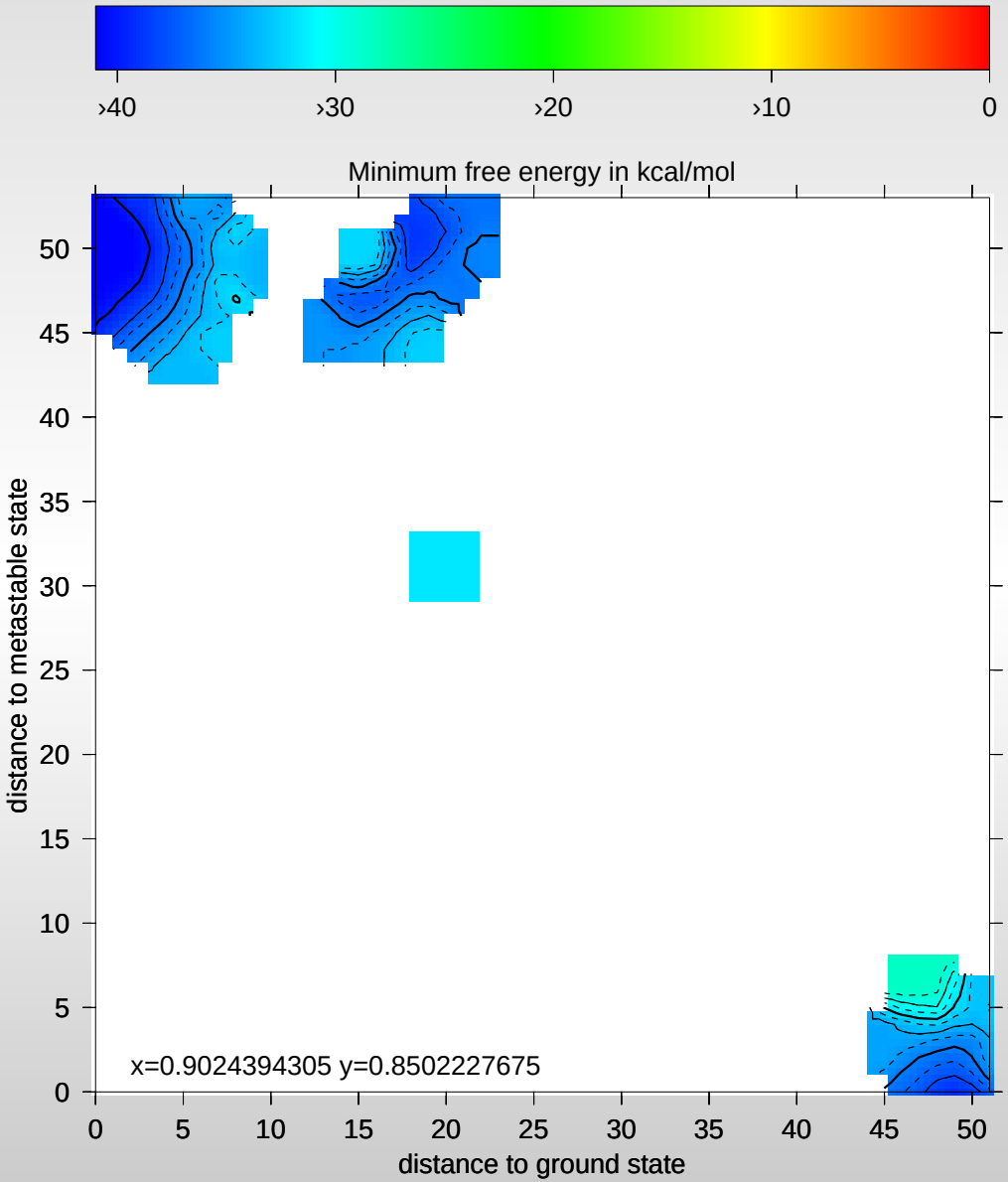
Example 1: designed RNA switch - distortion 10^4



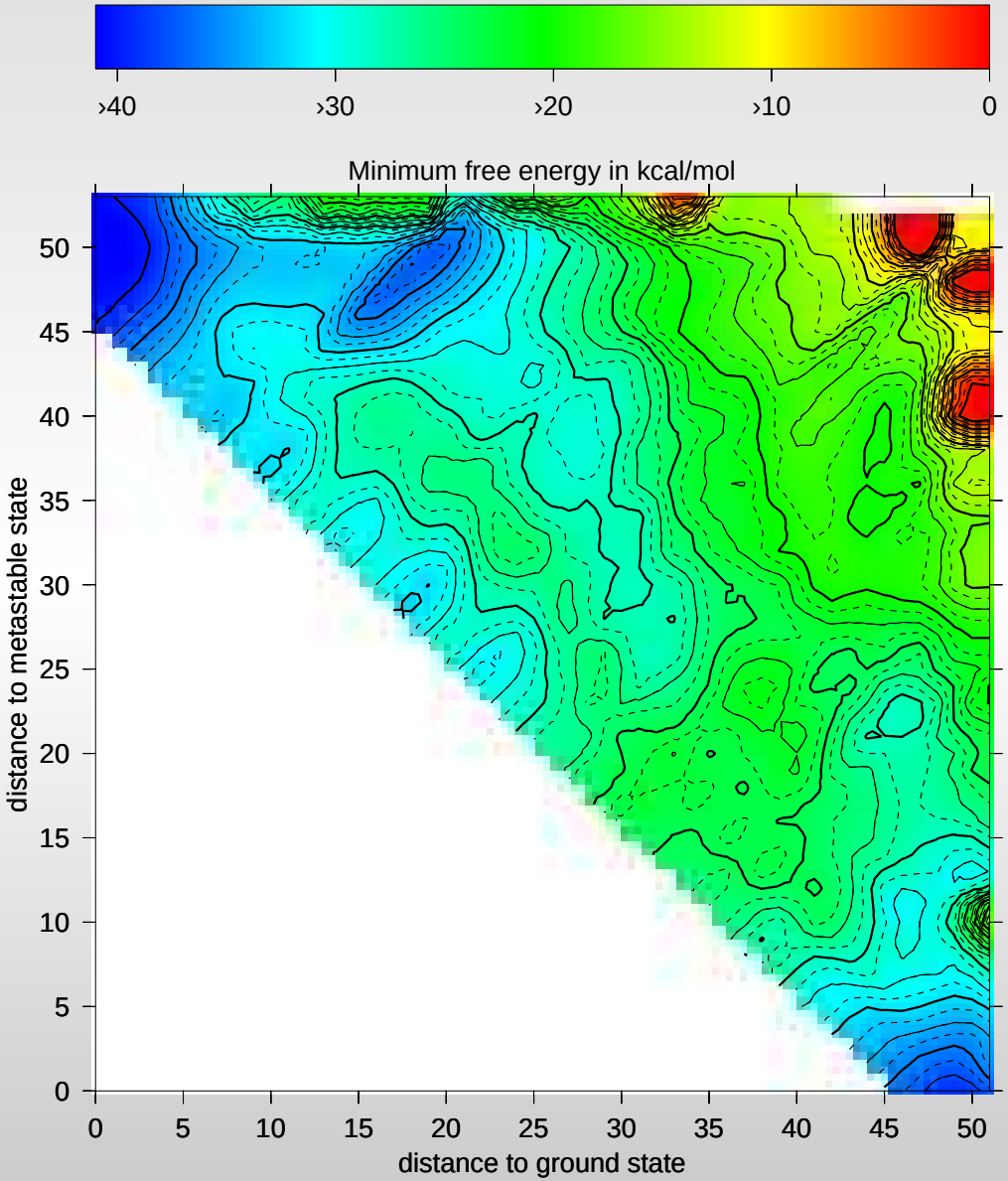
Example 1: designed RNA switch - distortion 10^5



Example 1: designed RNA switch - distortion 10^6

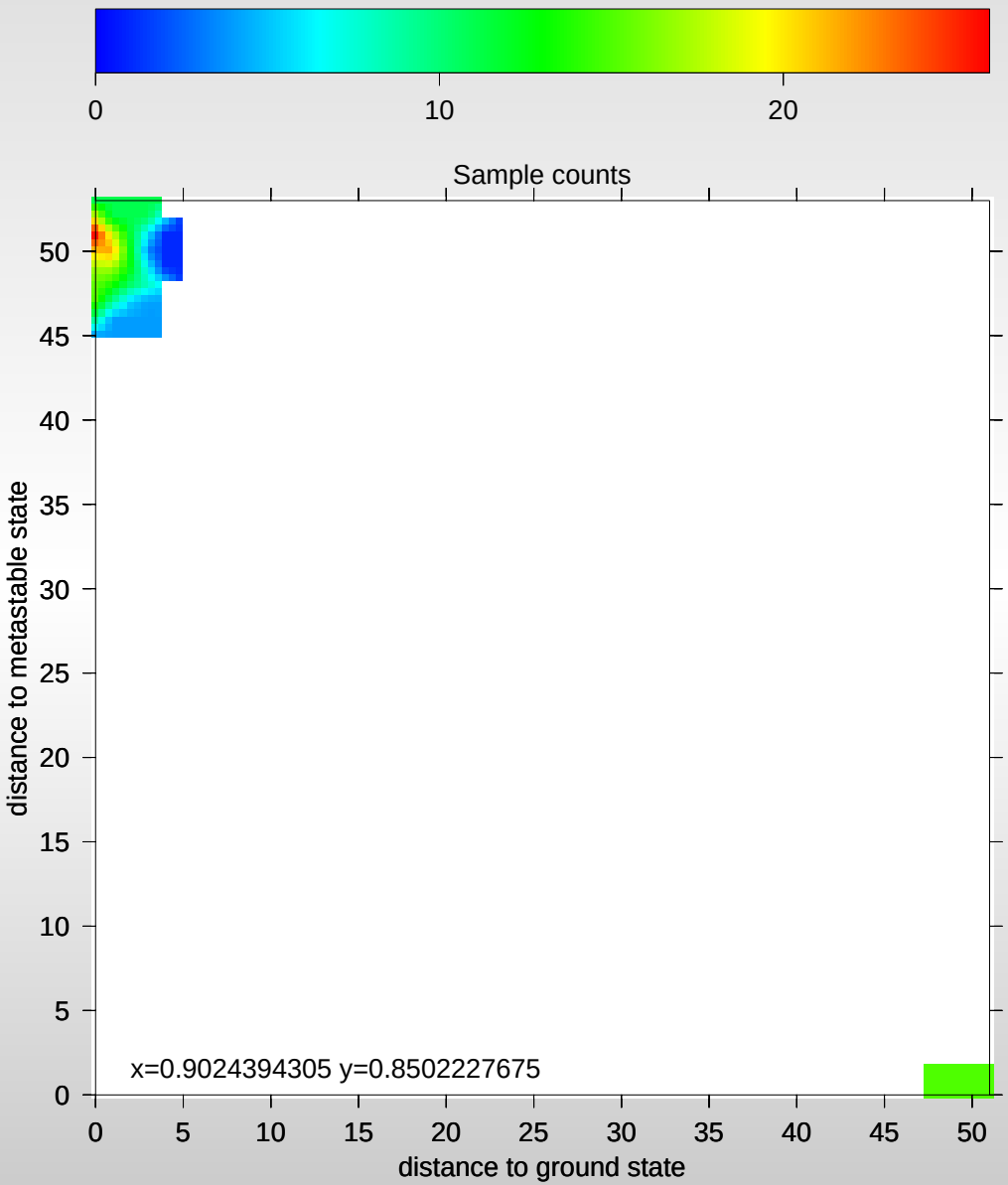


Example 1: designed RNA switch [9] - RNA2Dfold

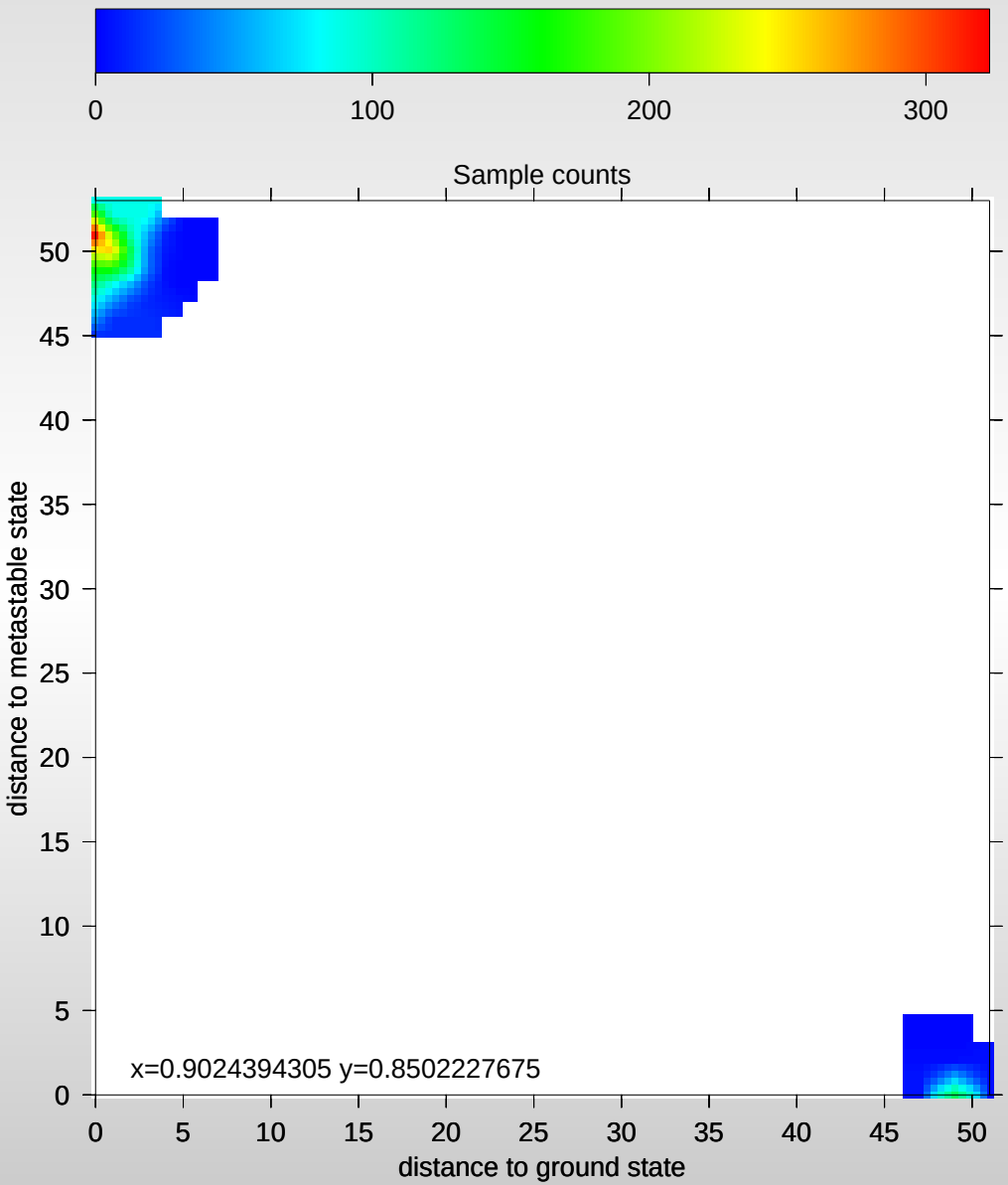


[9] Xayaphoummine et al. 2007

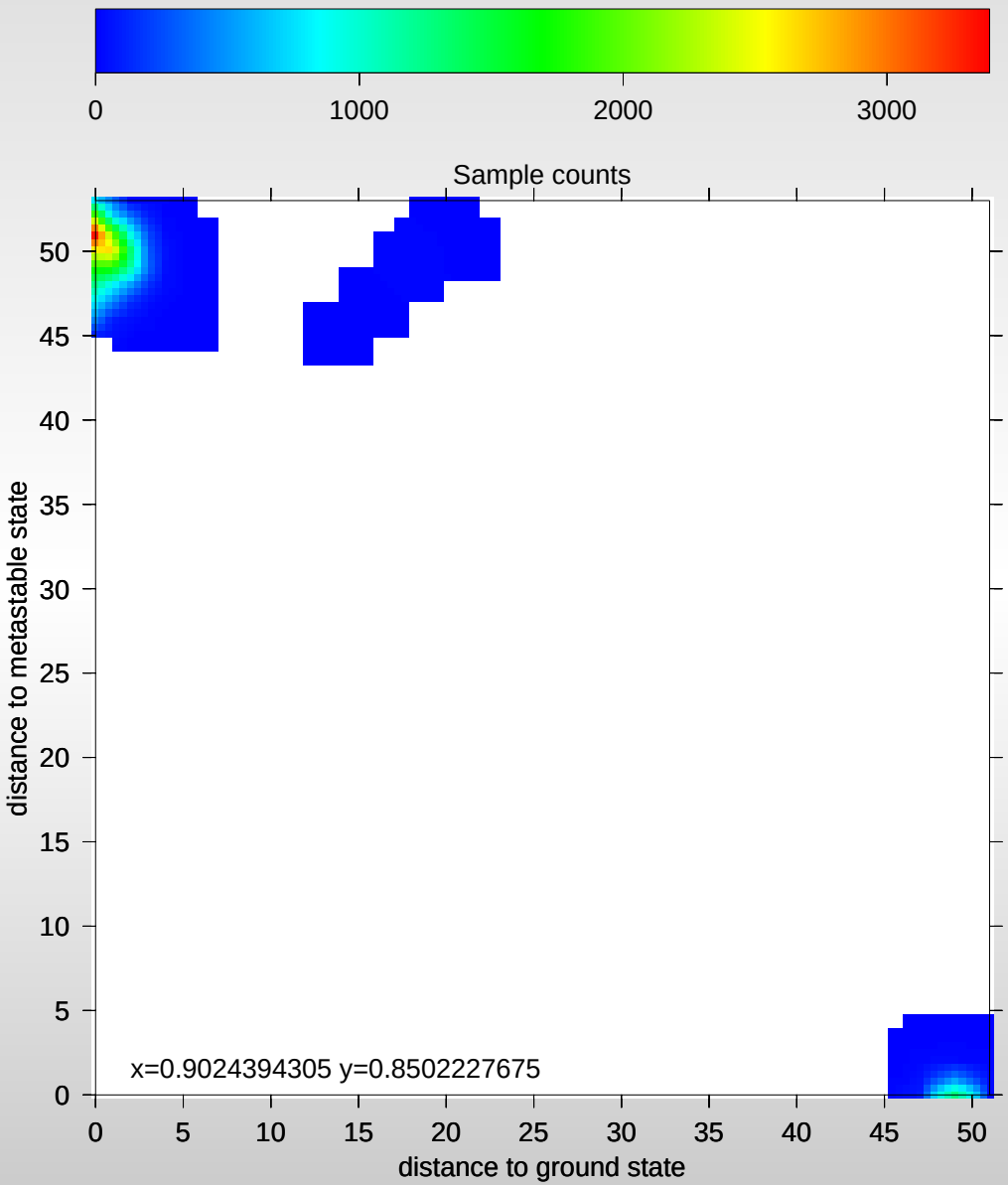
Example 1: designed RNA switch - distortion 10^2



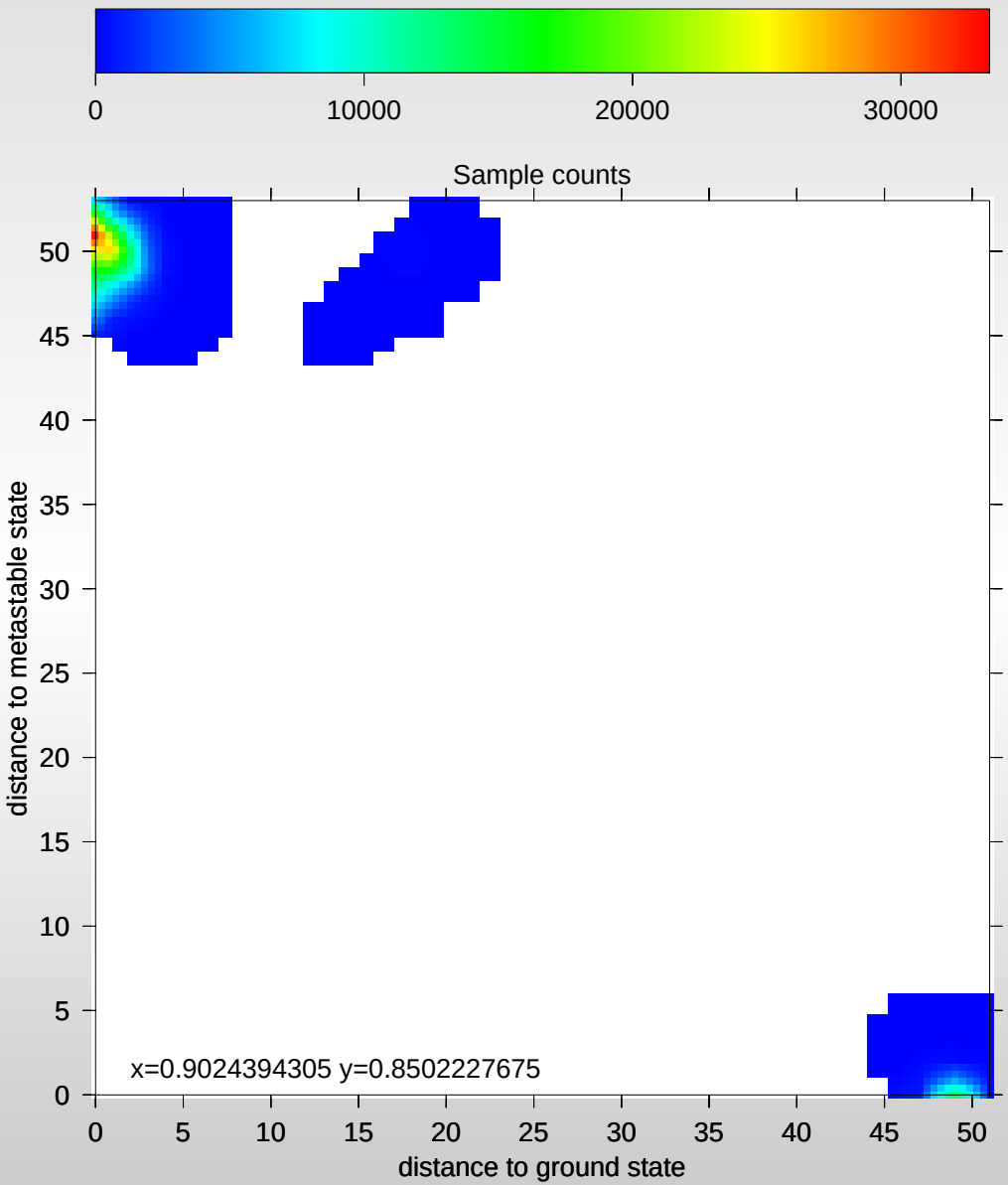
Example 1: designed RNA switch - distortion 10^3



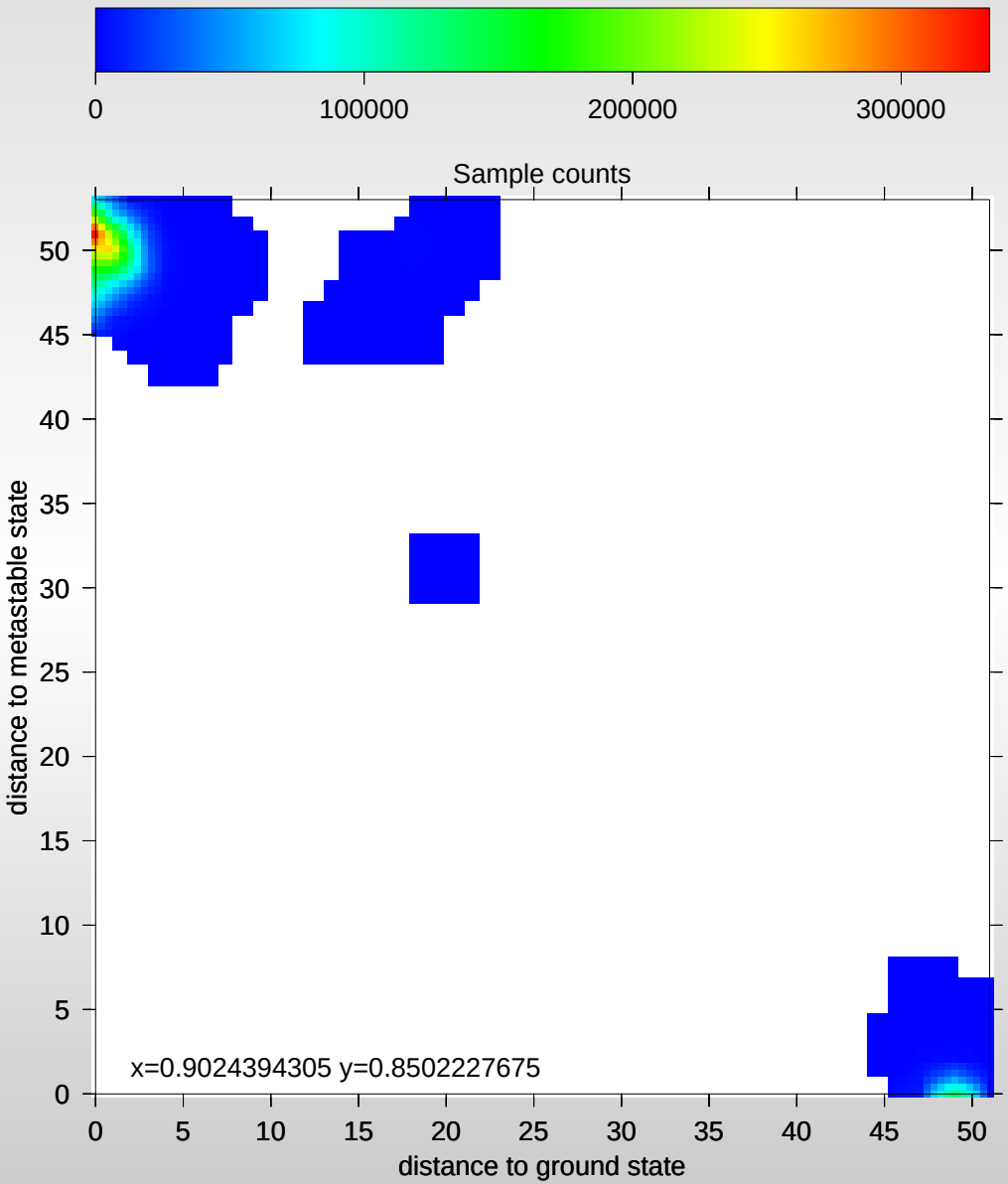
Example 1: designed RNA switch - distortion 10^4



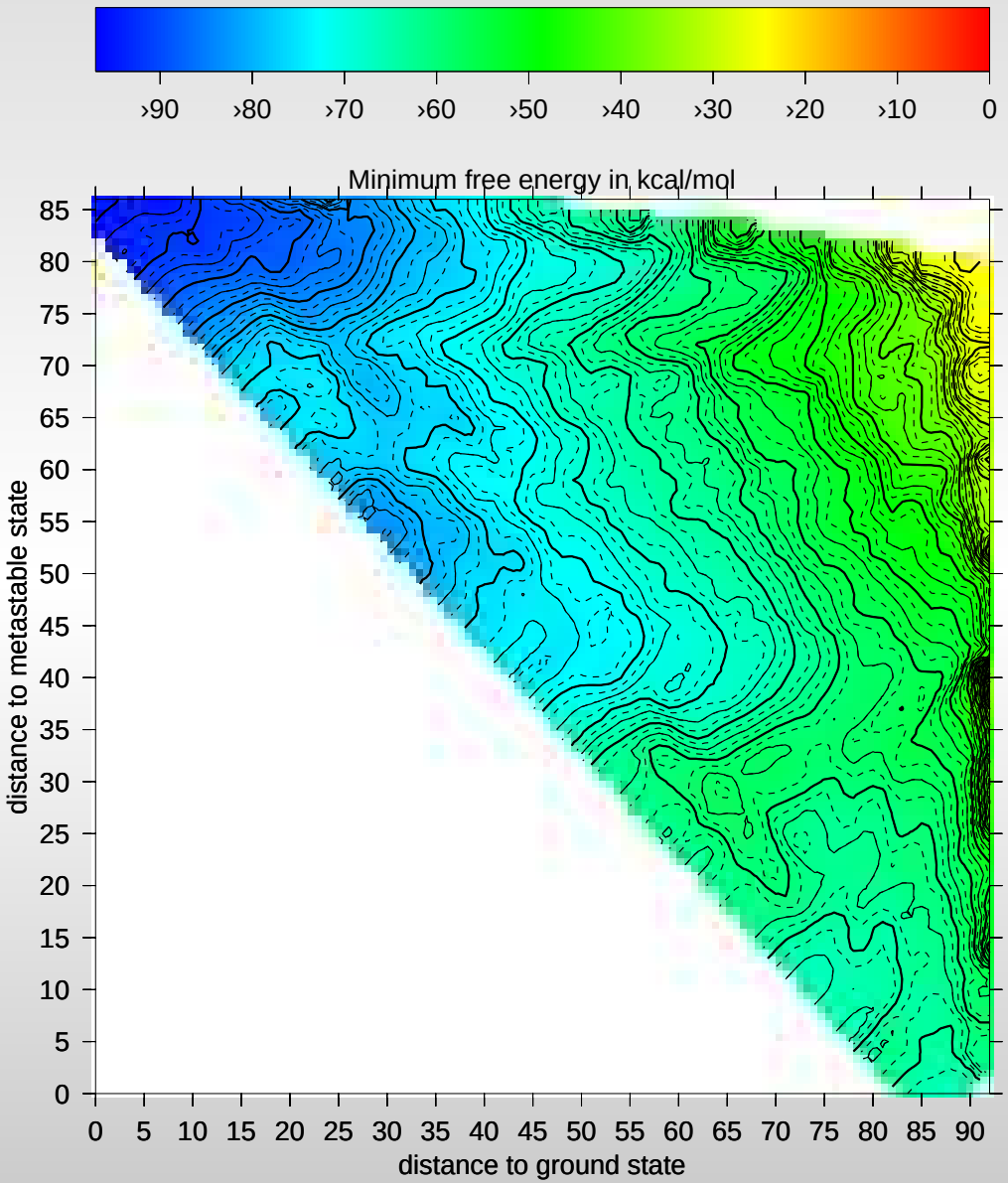
Example 1: designed RNA switch - distortion 10^5



Example 1: designed RNA switch - distortion 10^6

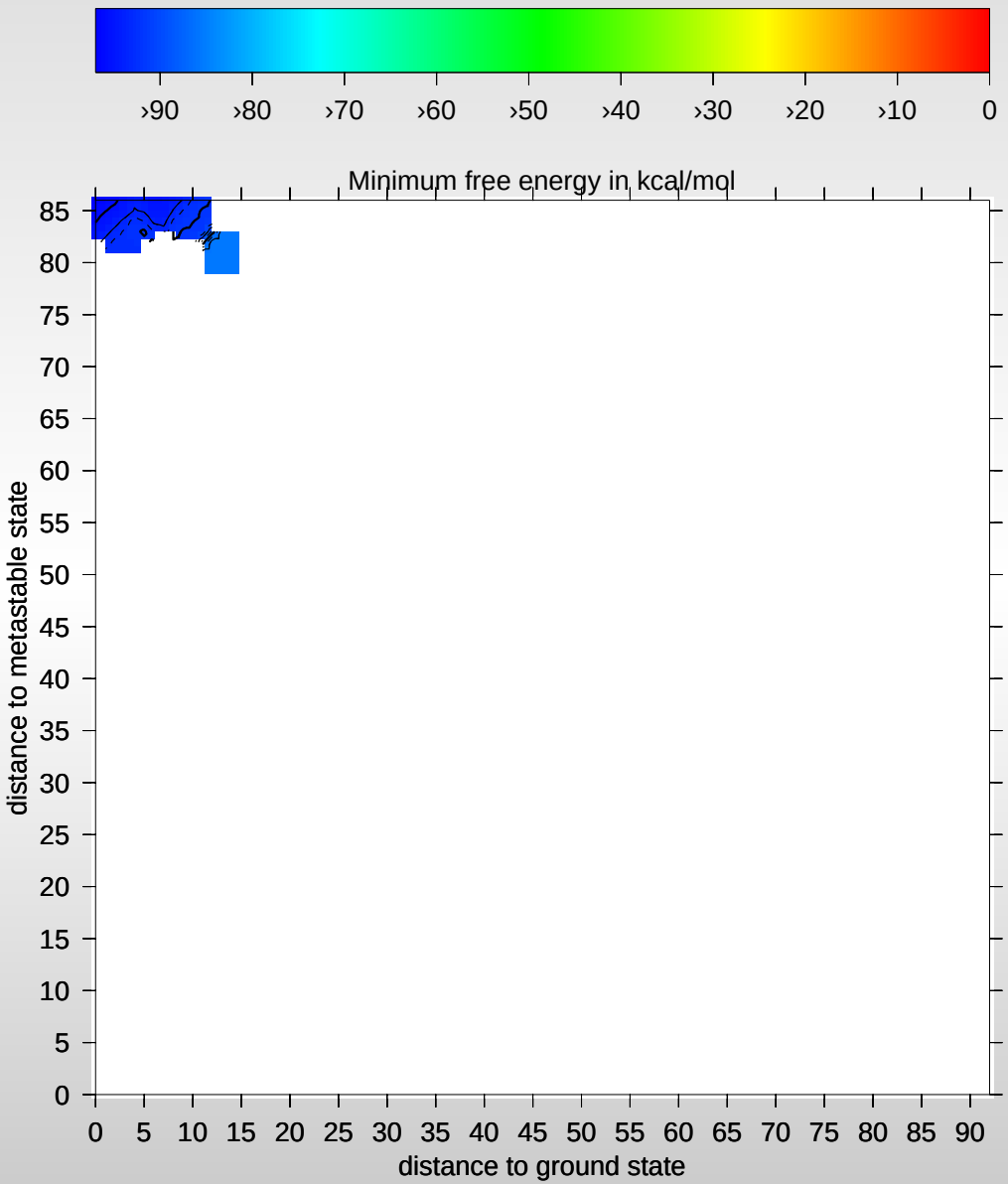


Example 2: SV11 RNA [10] - RNA2Dfold

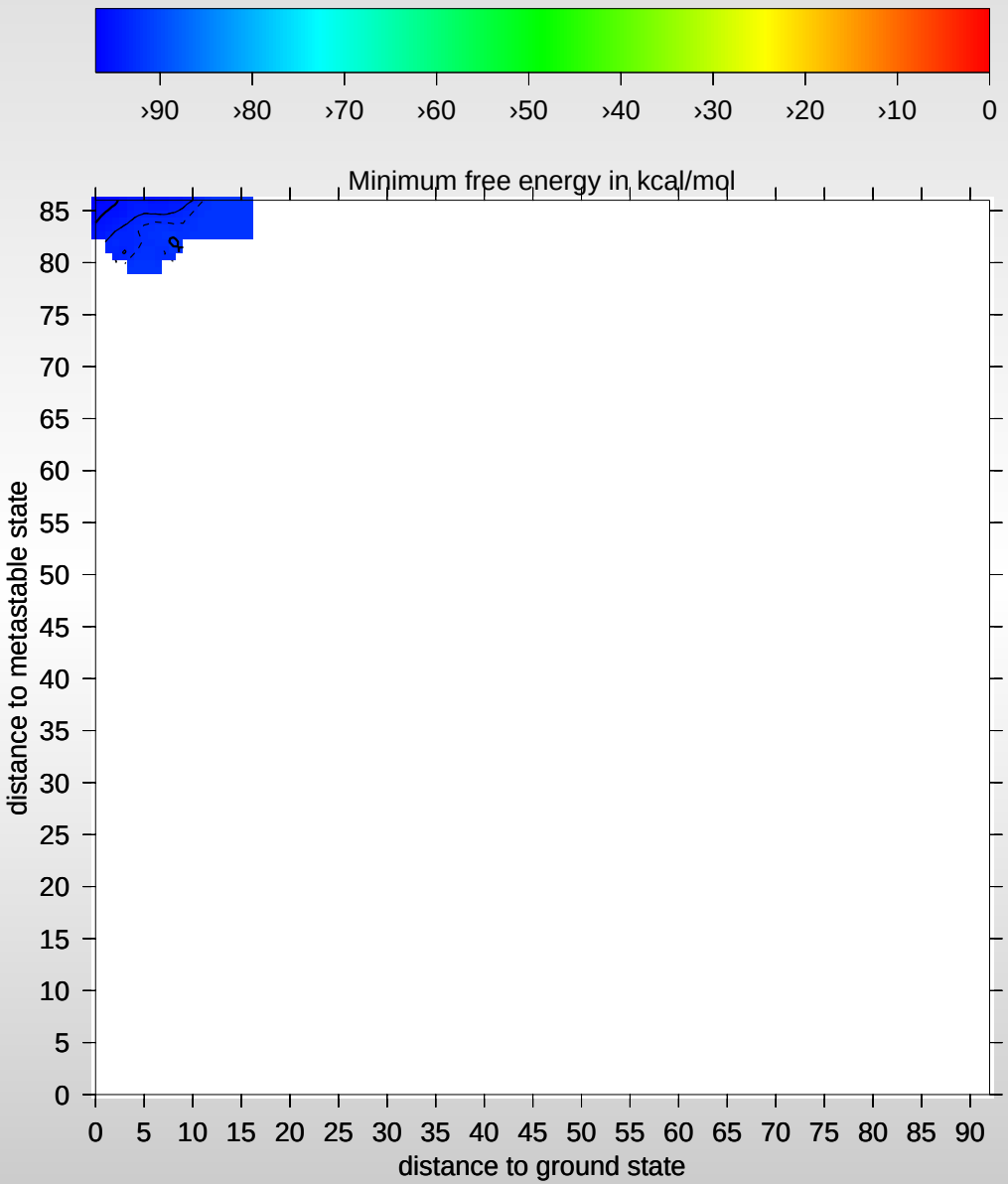


[10] Biebricher et al. 1982, Biebricher and Luce 1992

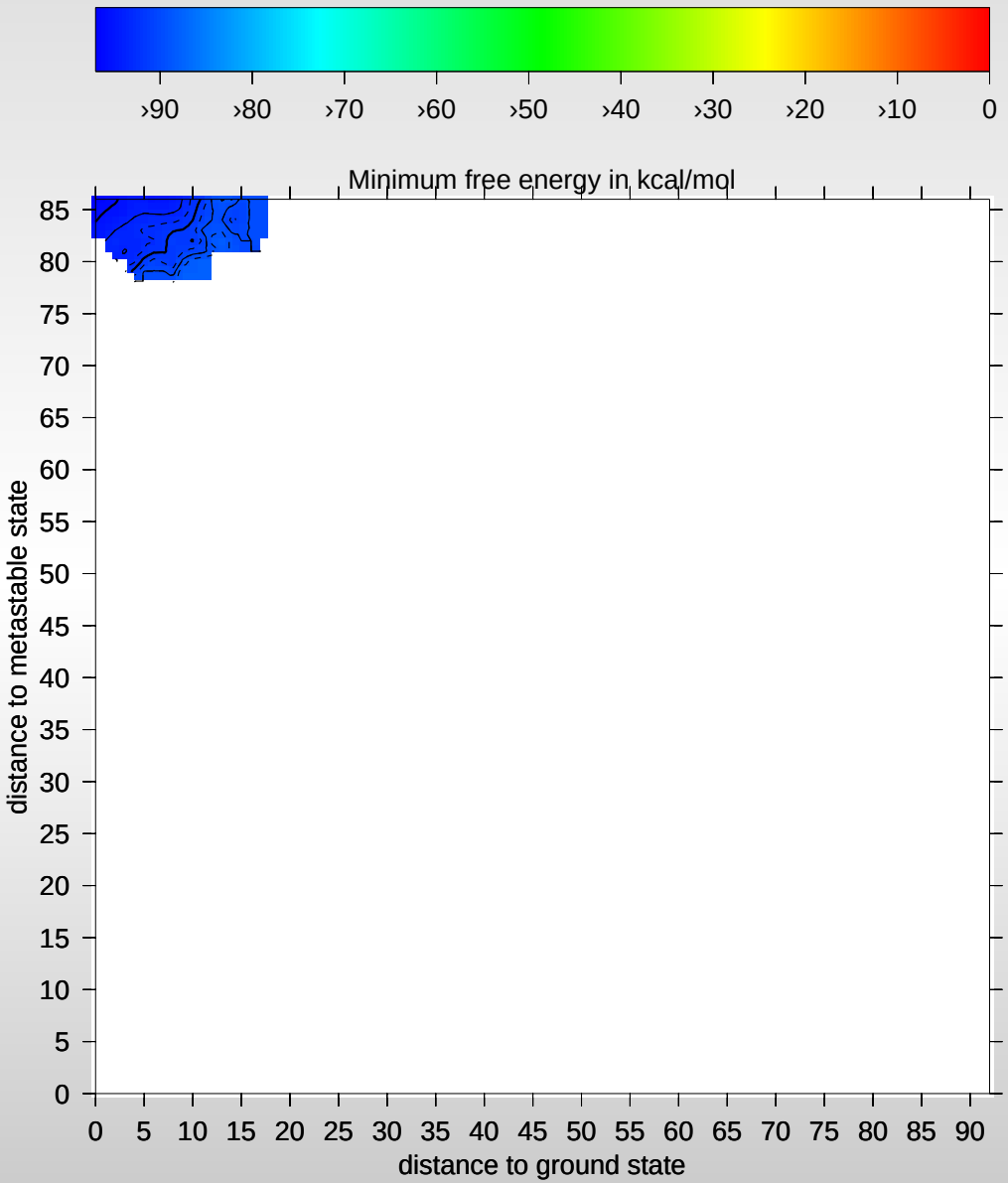
Example 2: SV11 RNA - RNAsubopt -p 10^2



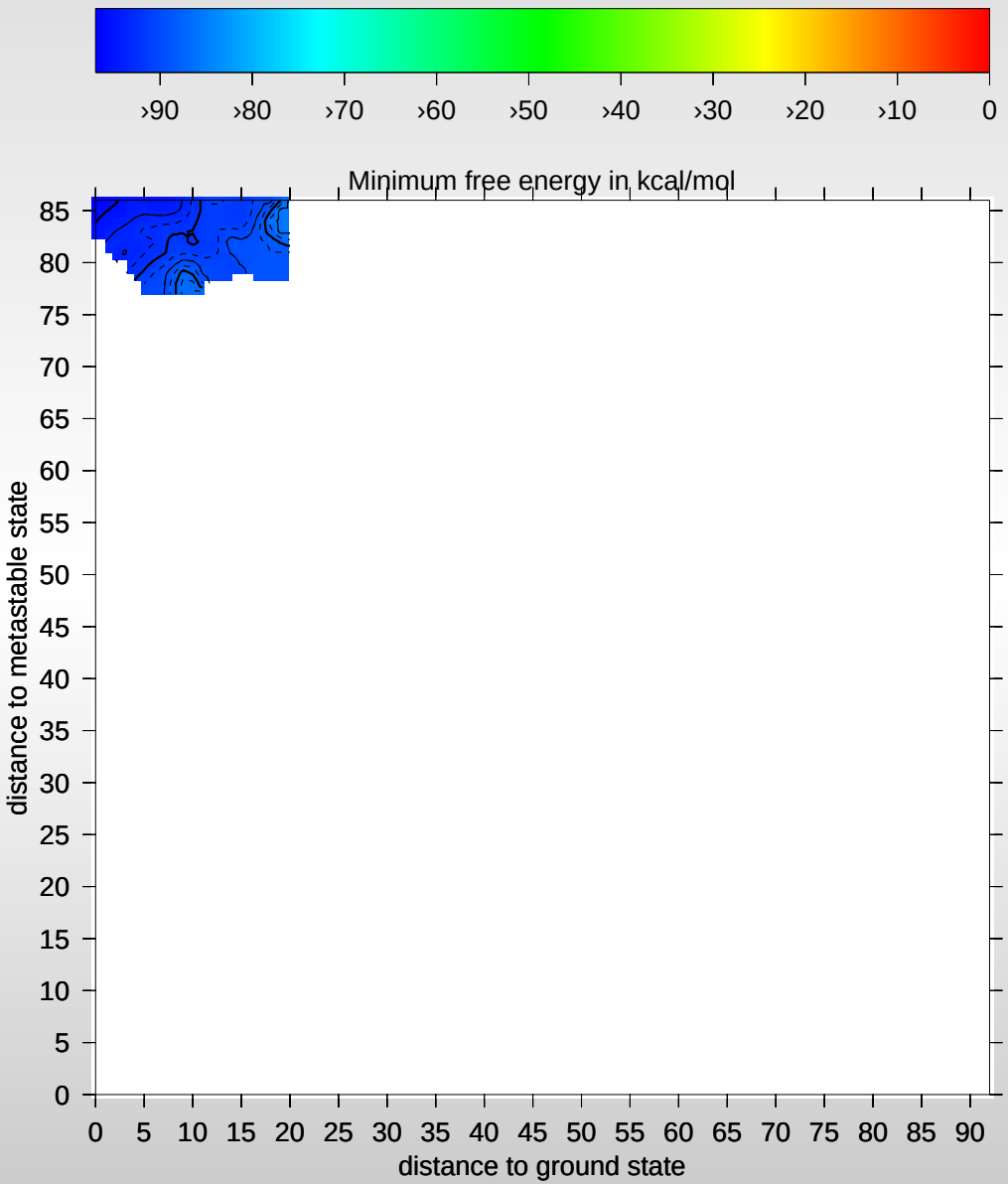
Example 2: SV11 RNA - RNAsubopt -p 10^3



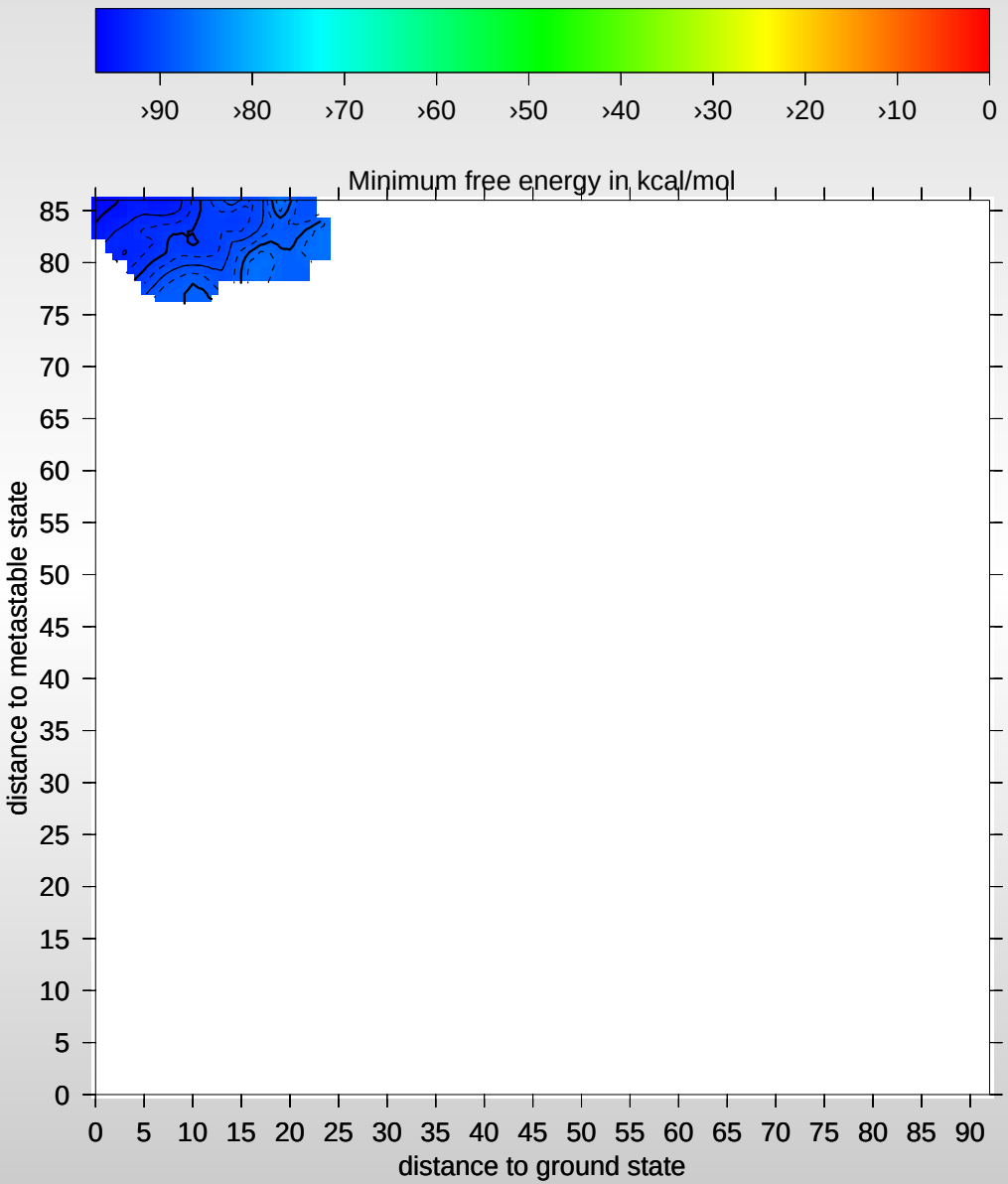
Example 2: SV11 RNA - RNAsubopt -p 10^4



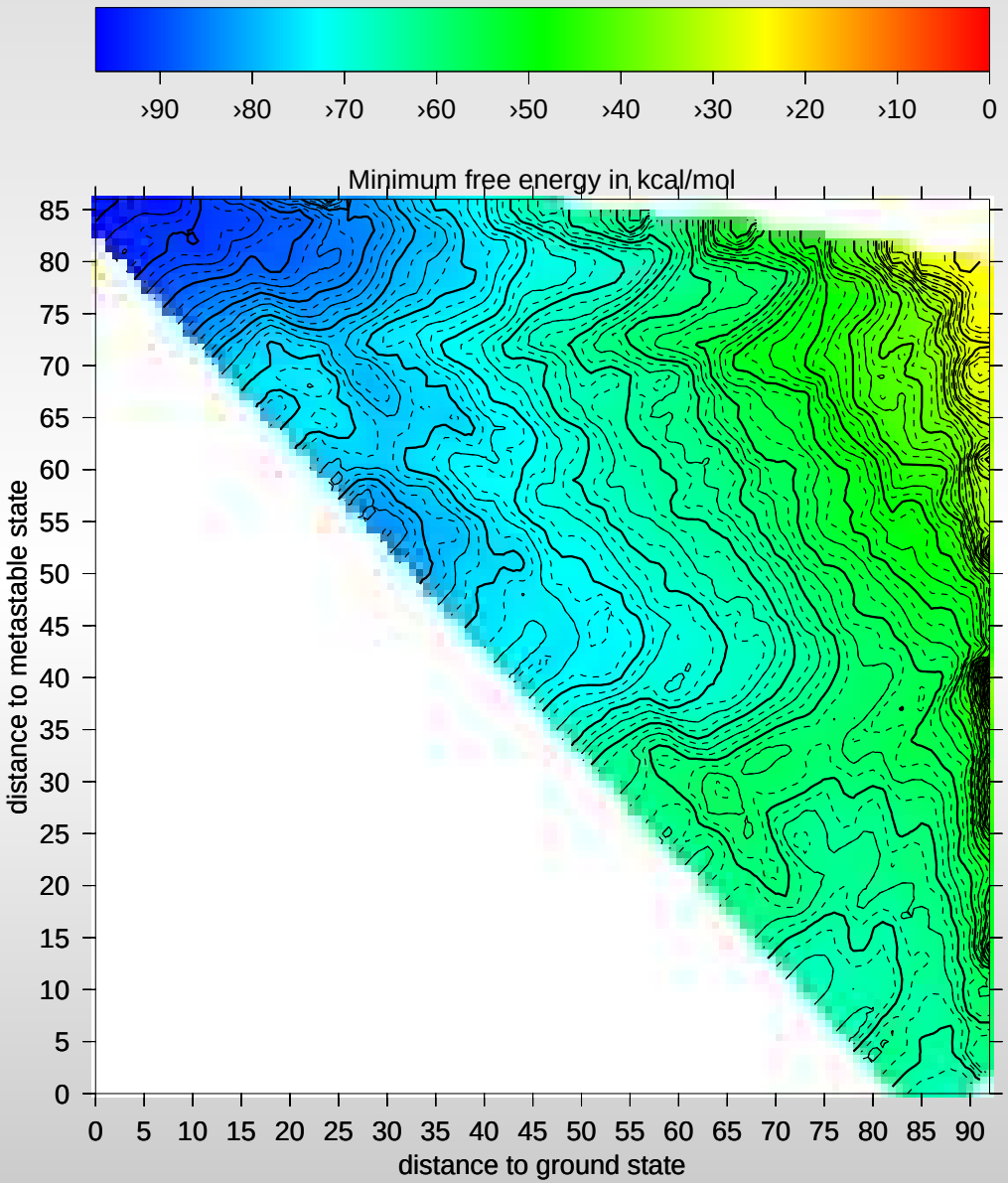
Example 2: SV11 RNA - RNAsubopt -p 10^5



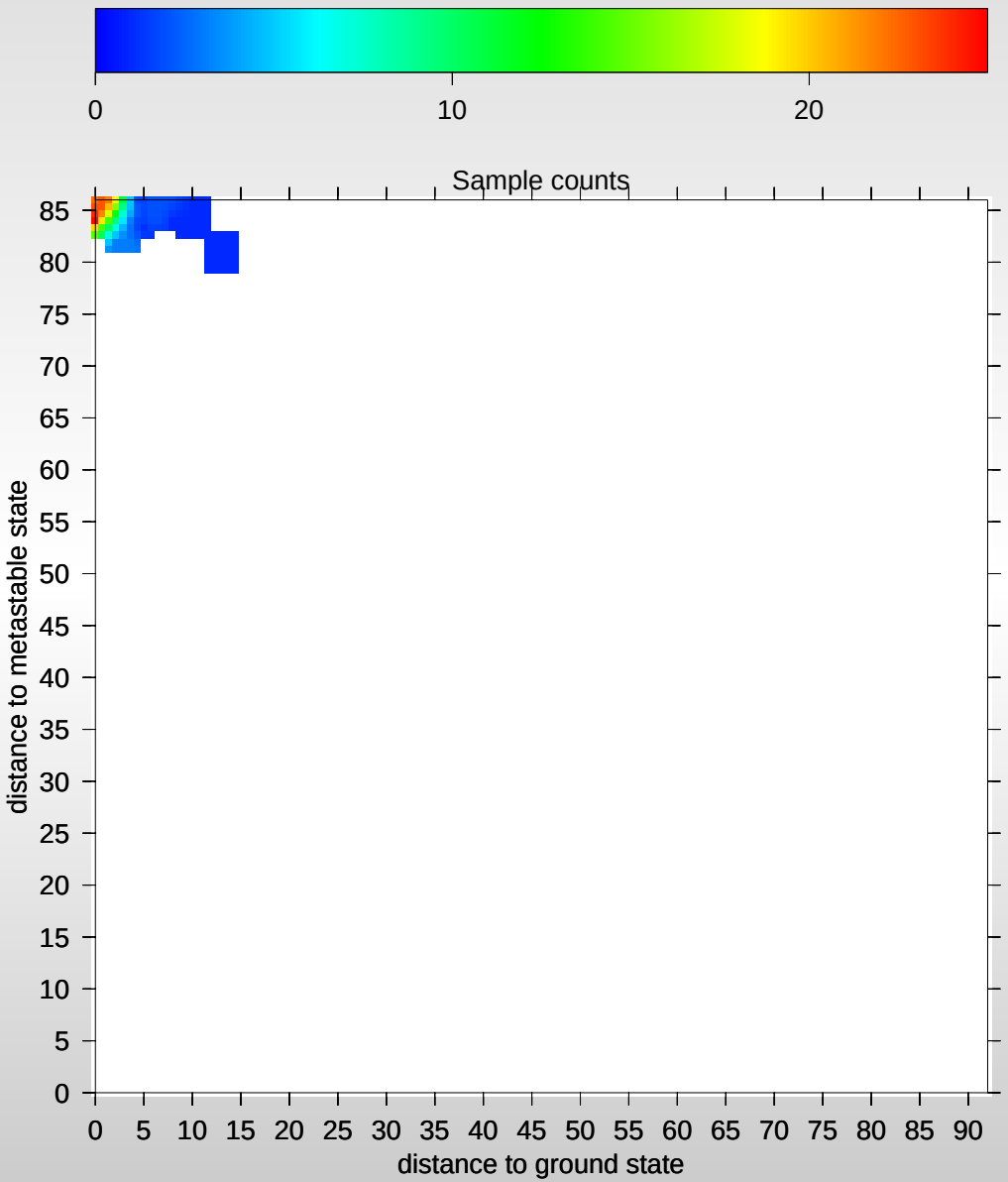
Example 2: SV11 RNA - RNAsubopt -p 10^6



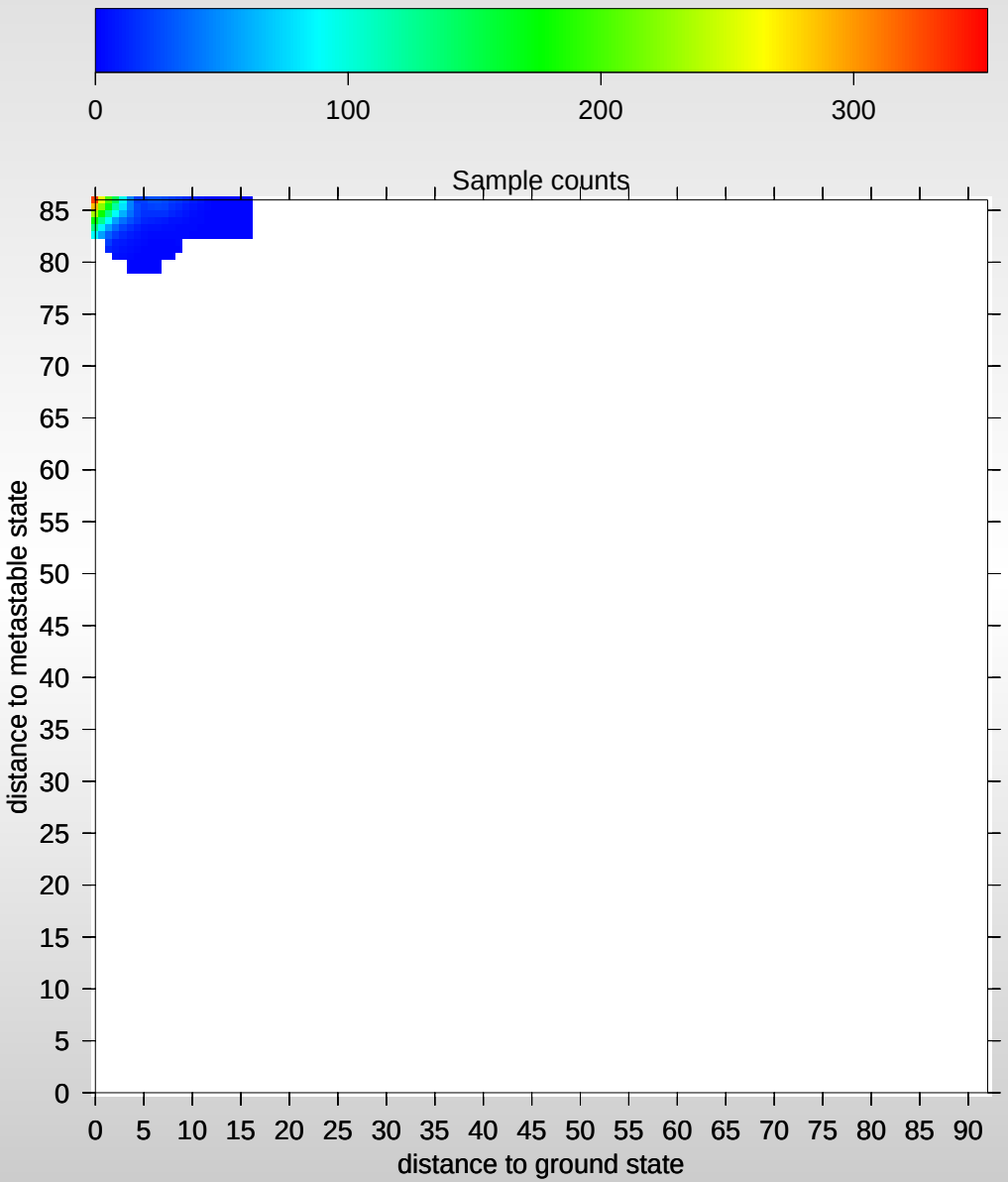
Example 2: SV11 RNA - RNA2Dfold



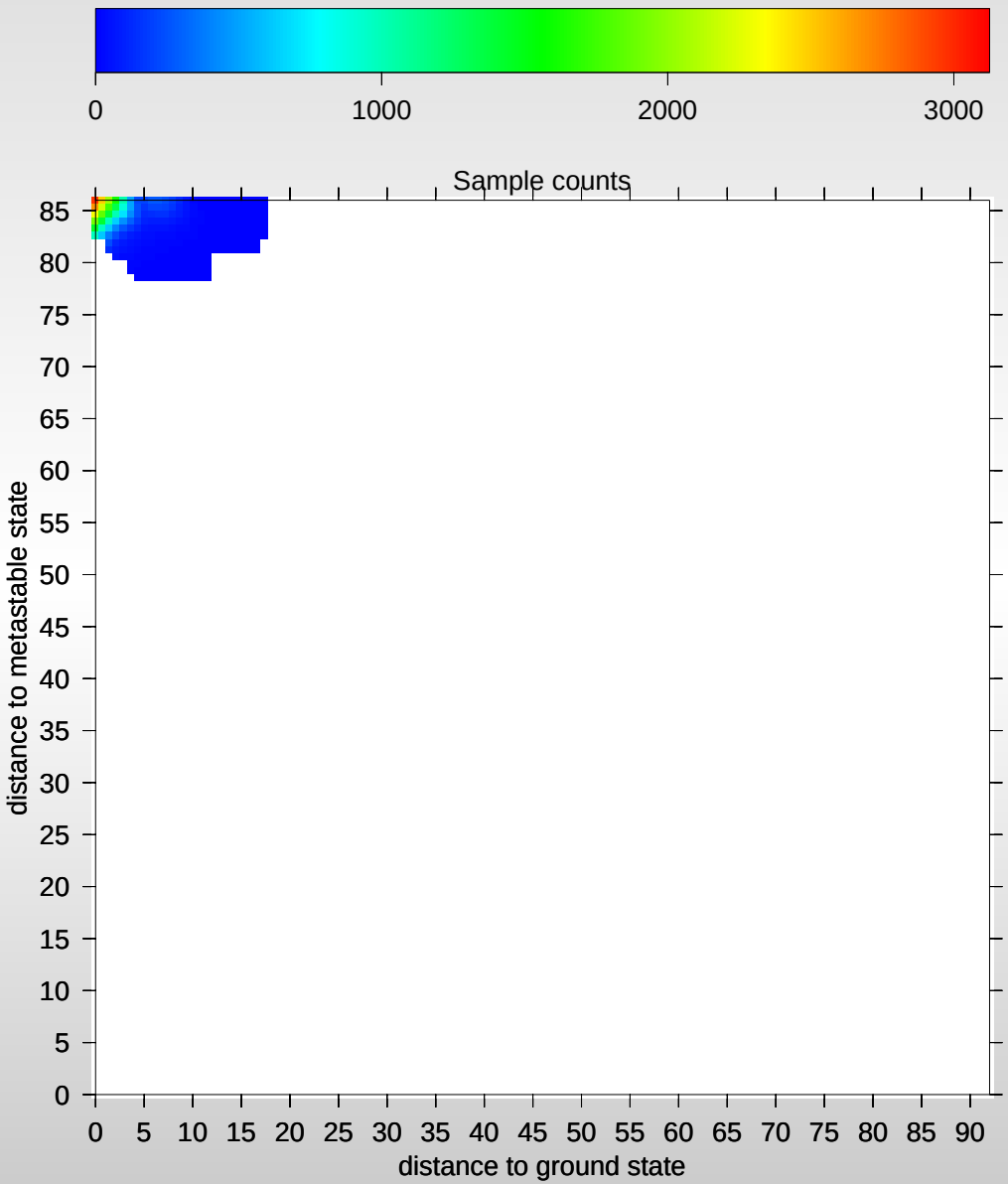
Example 2: SV11 RNA - RNAsubopt -p 10^2



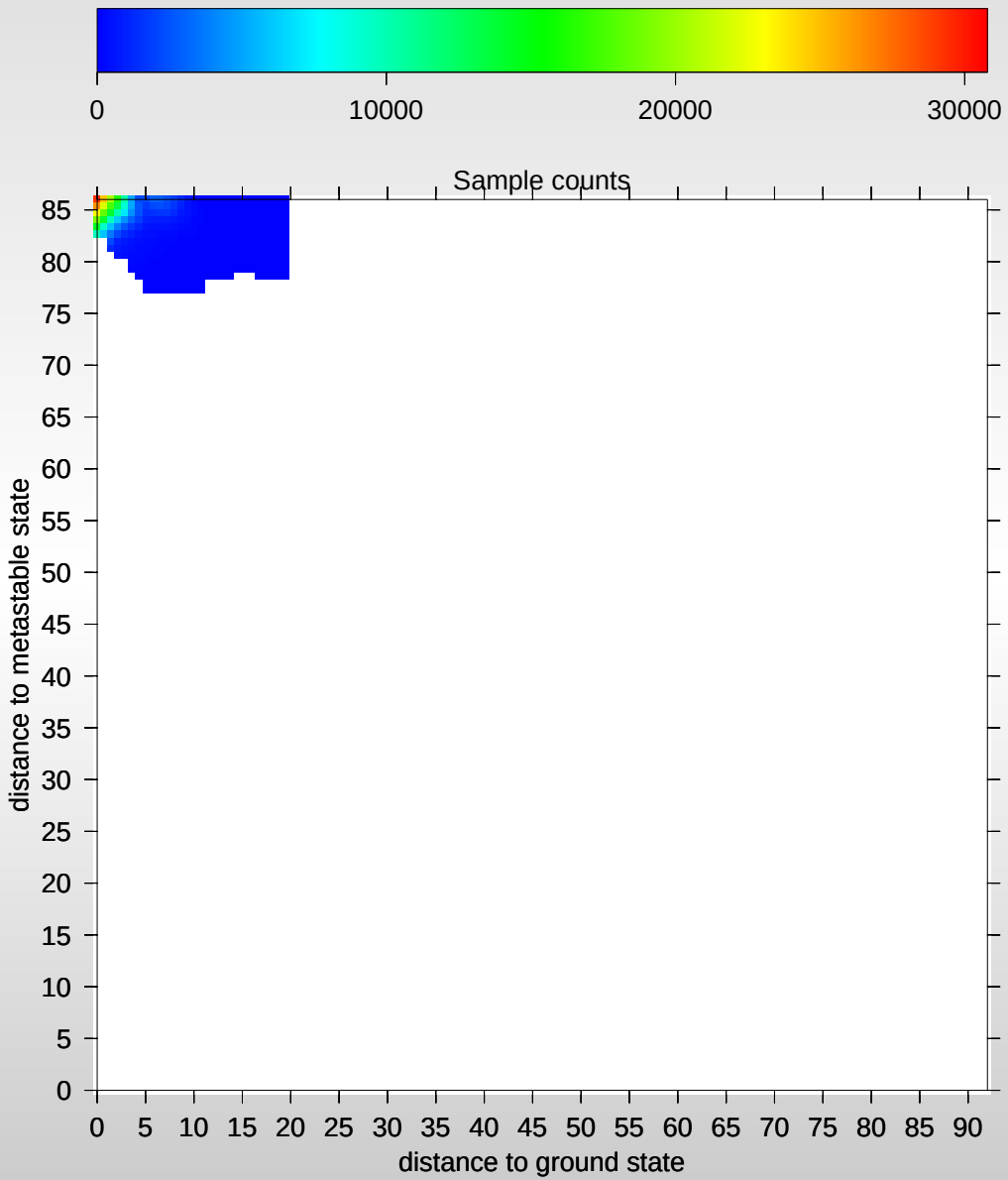
Example 2: SV11 RNA - RNAsubopt -p 10^3



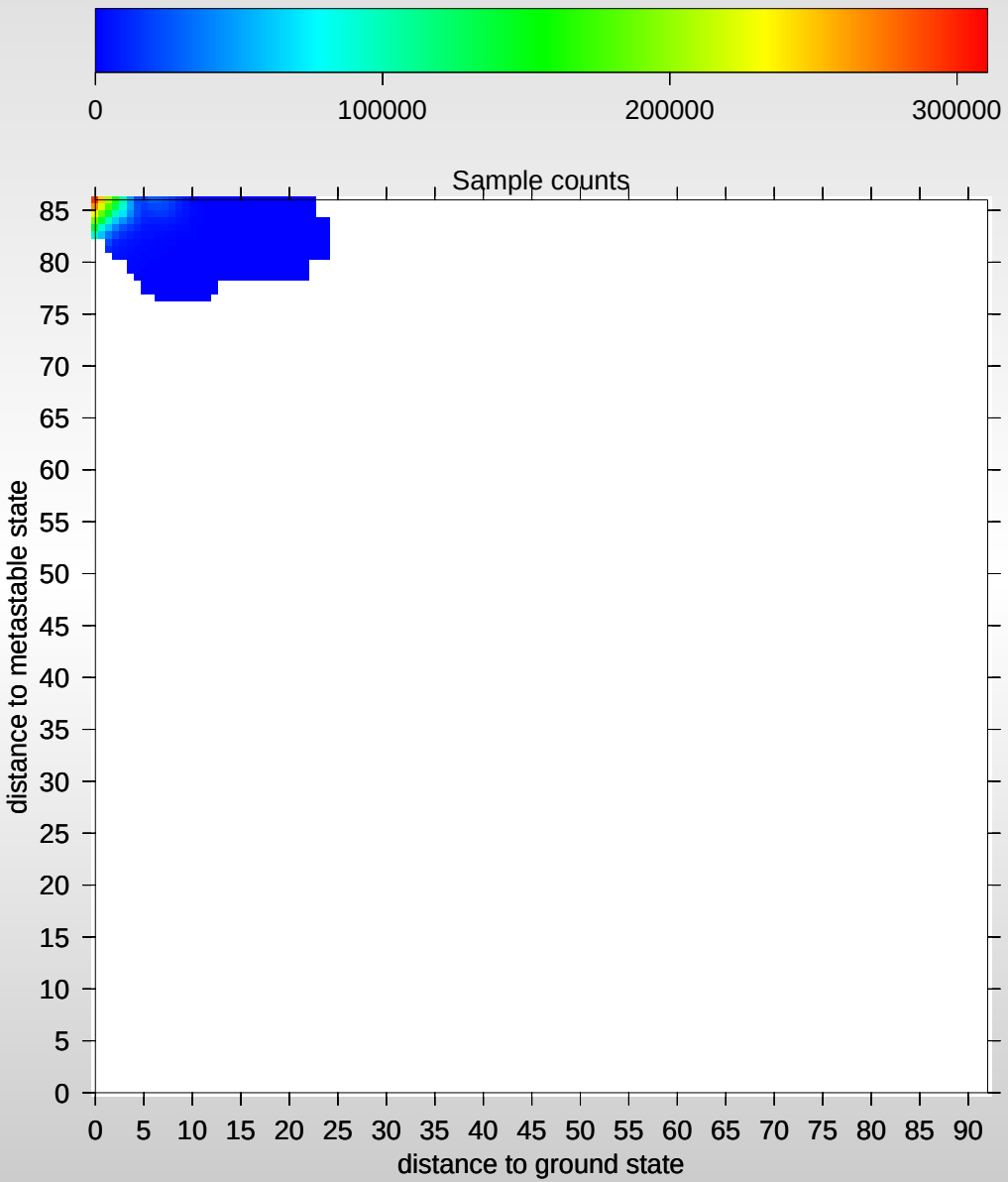
Example 2: SV11 RNA - RNAsubopt -p 10^4



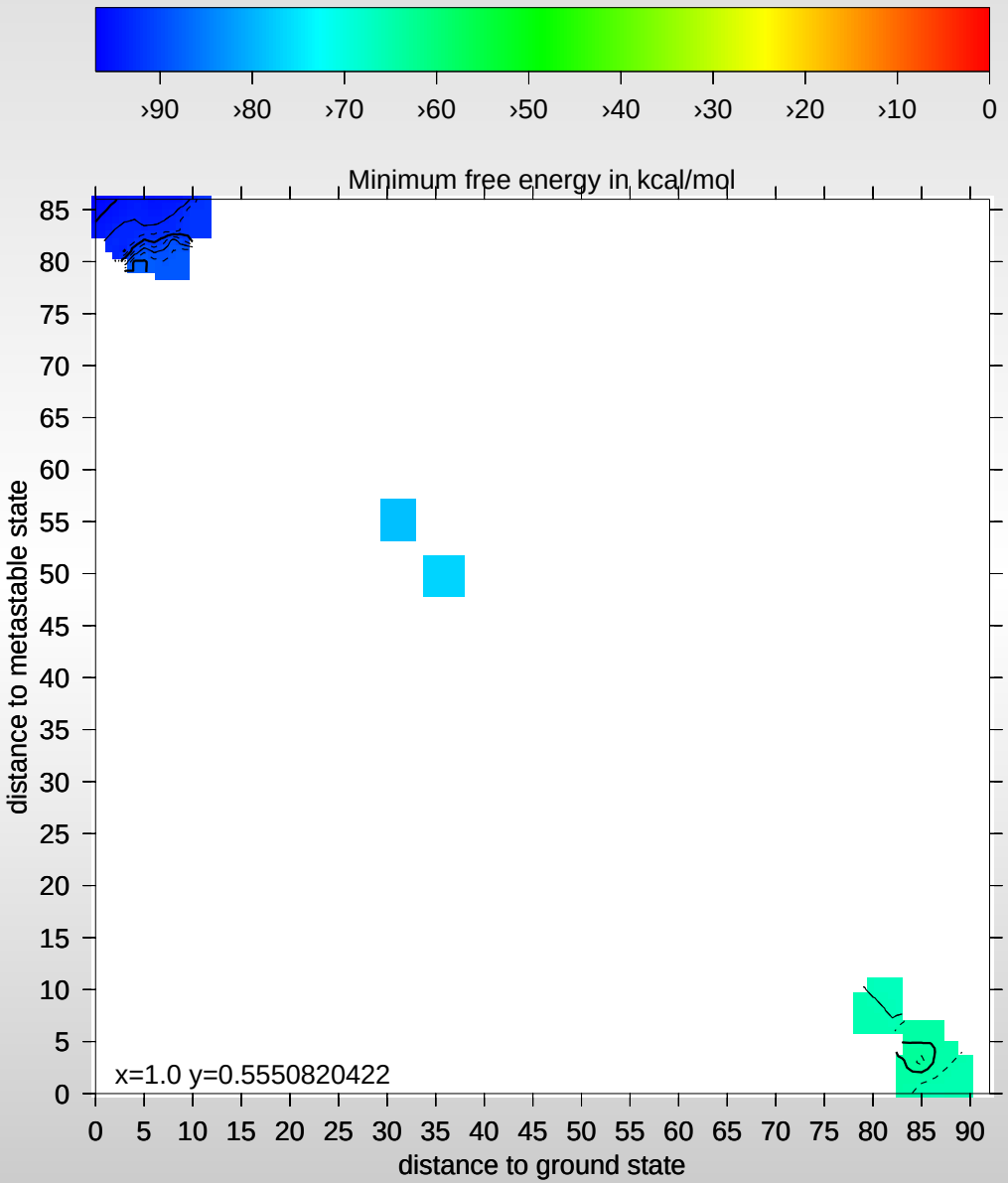
Example 2: SV11 RNA - RNAsubopt -p 10^5



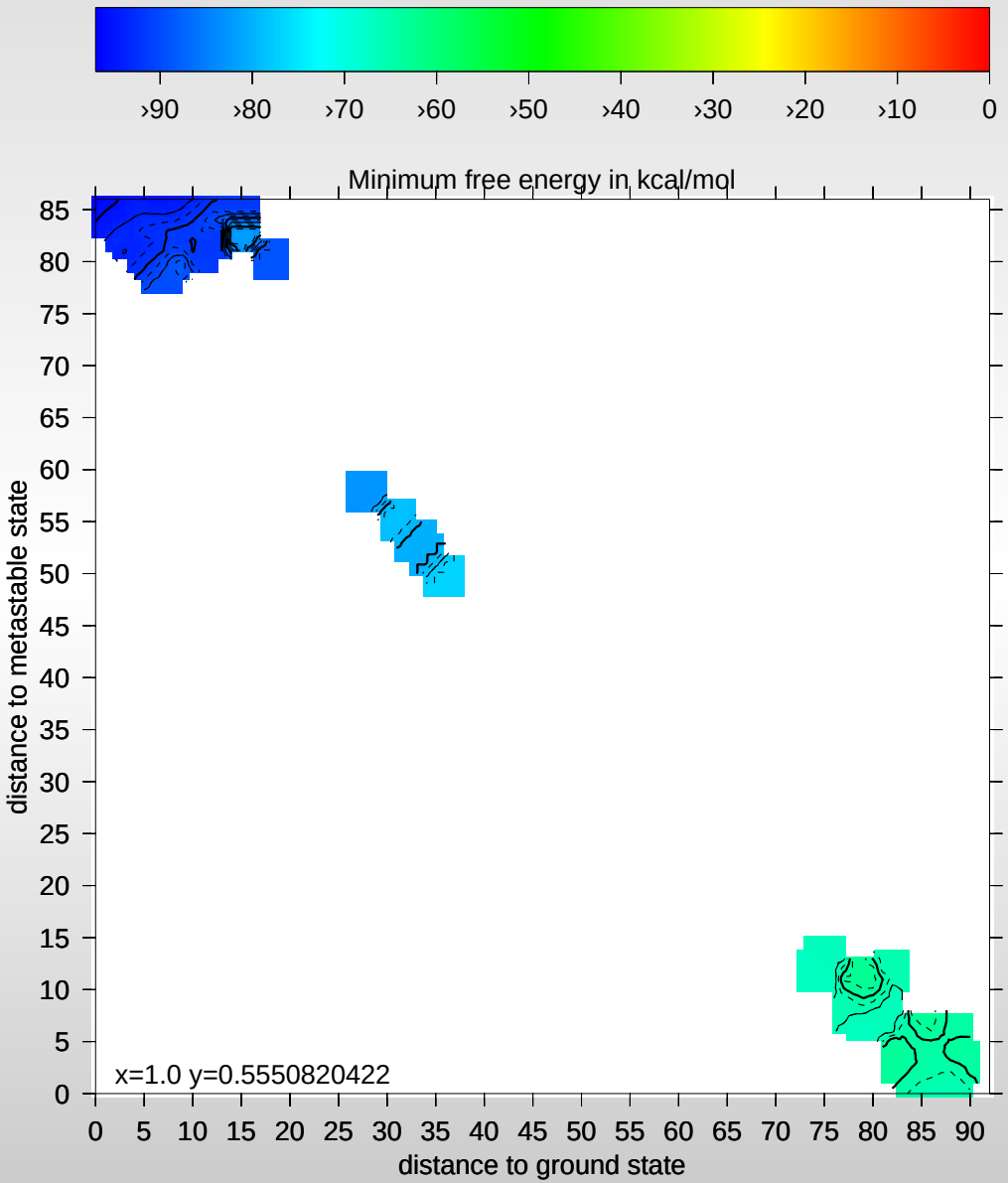
Example 2: SV11 RNA - RNAsubopt -p 10^6



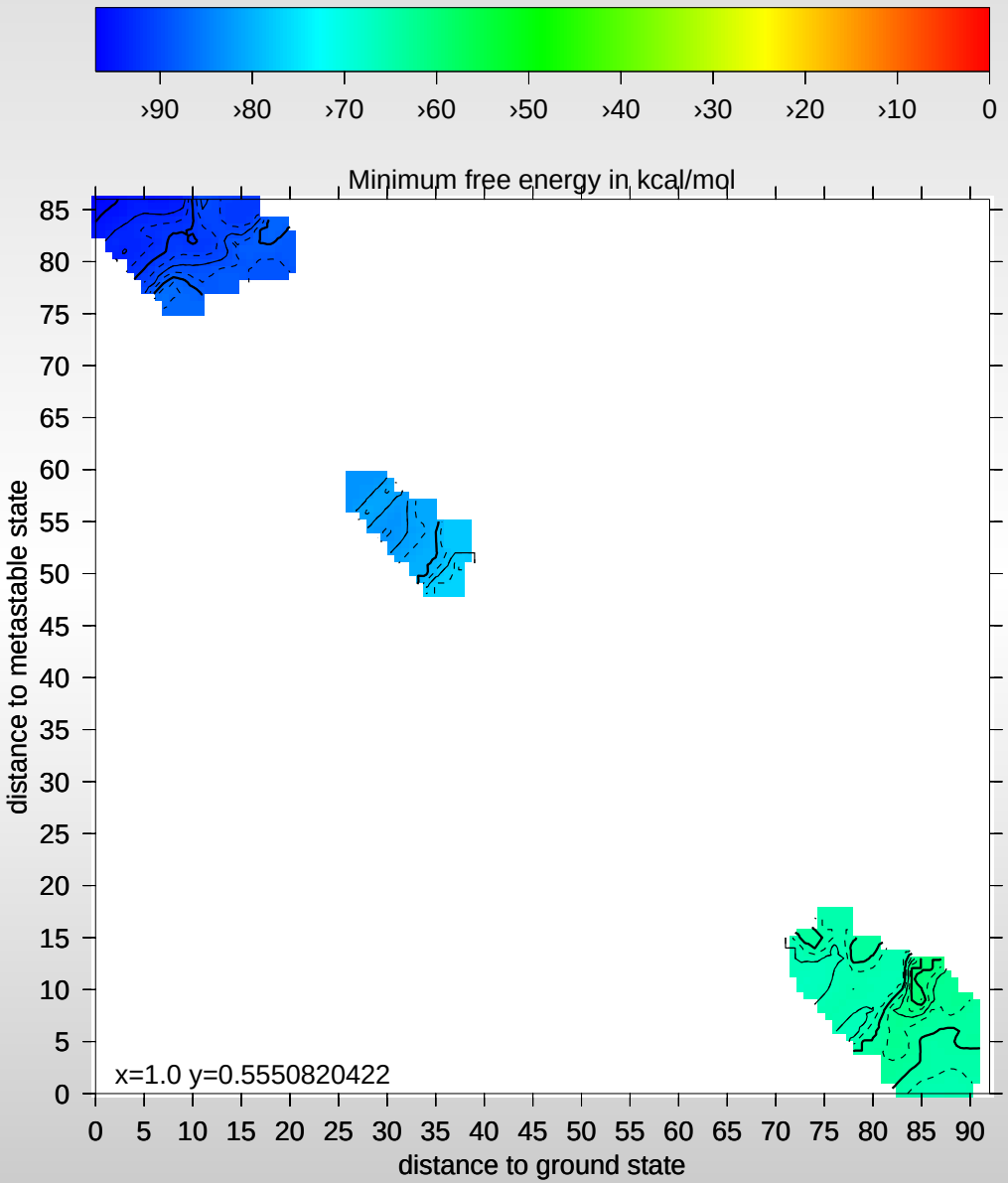
Example 2: SV11 RNA - distortion 10^2



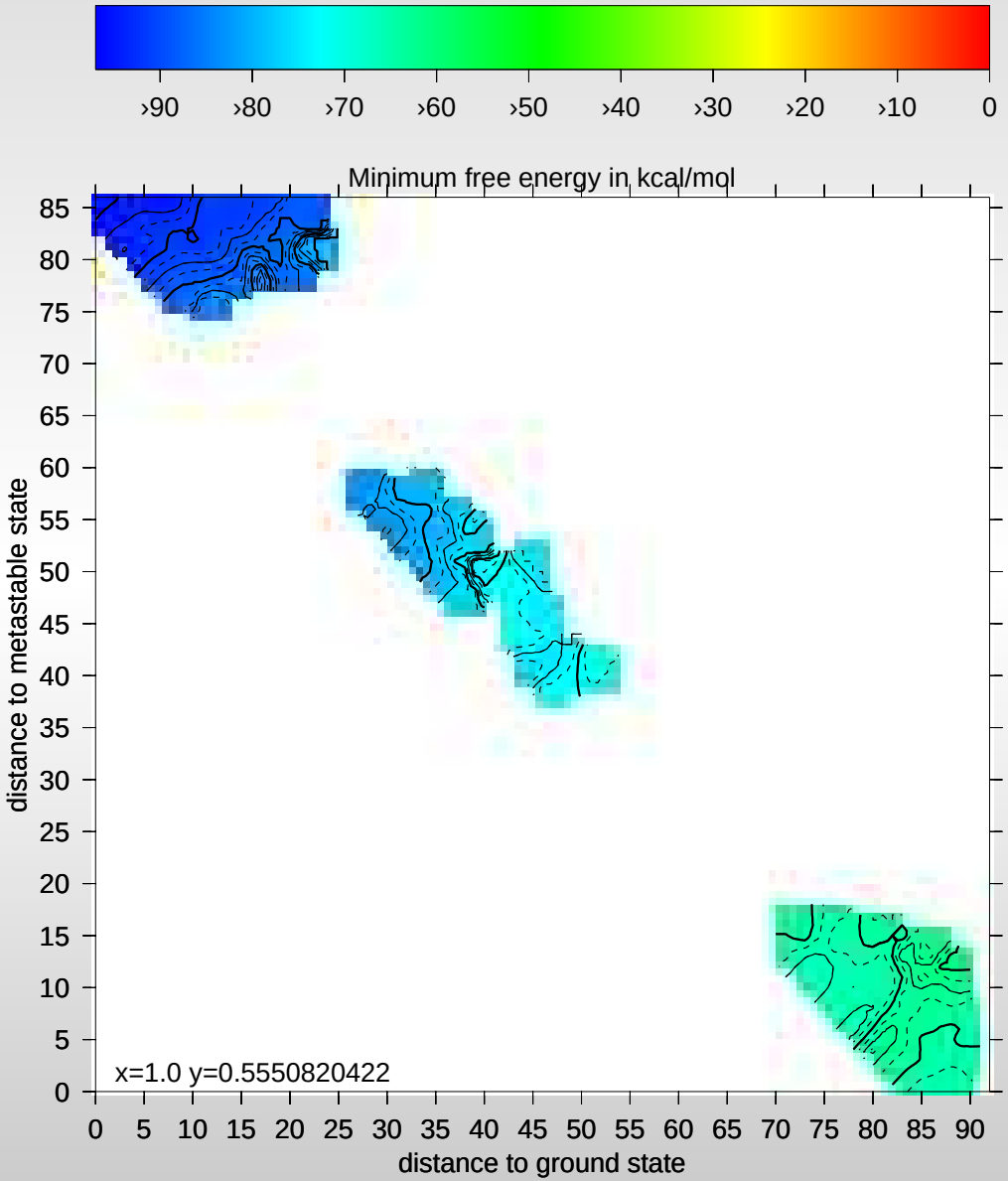
Example 2: SV11 RNA - distortion 10^3



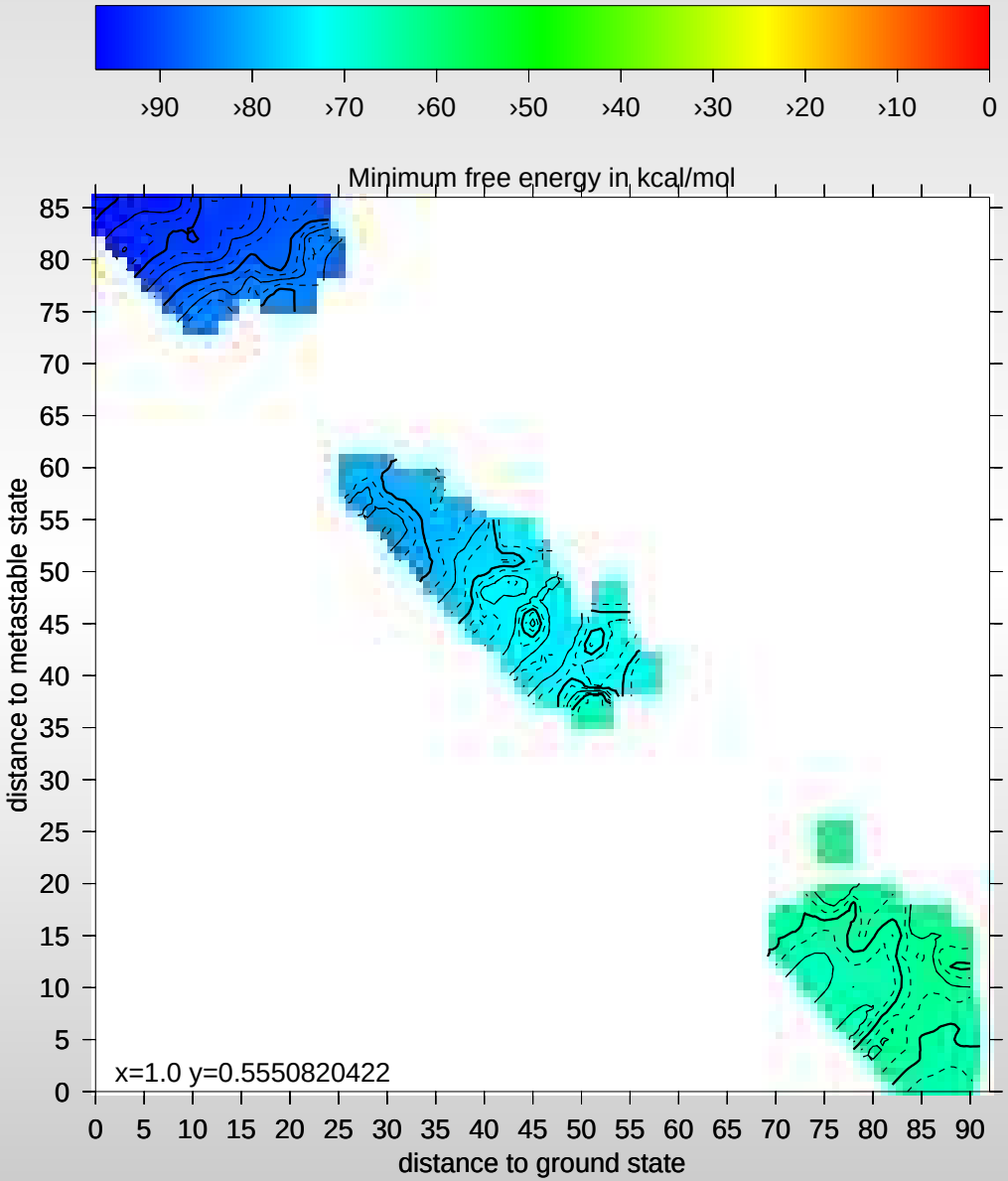
Example 2: SV11 RNA - distortion 10^4



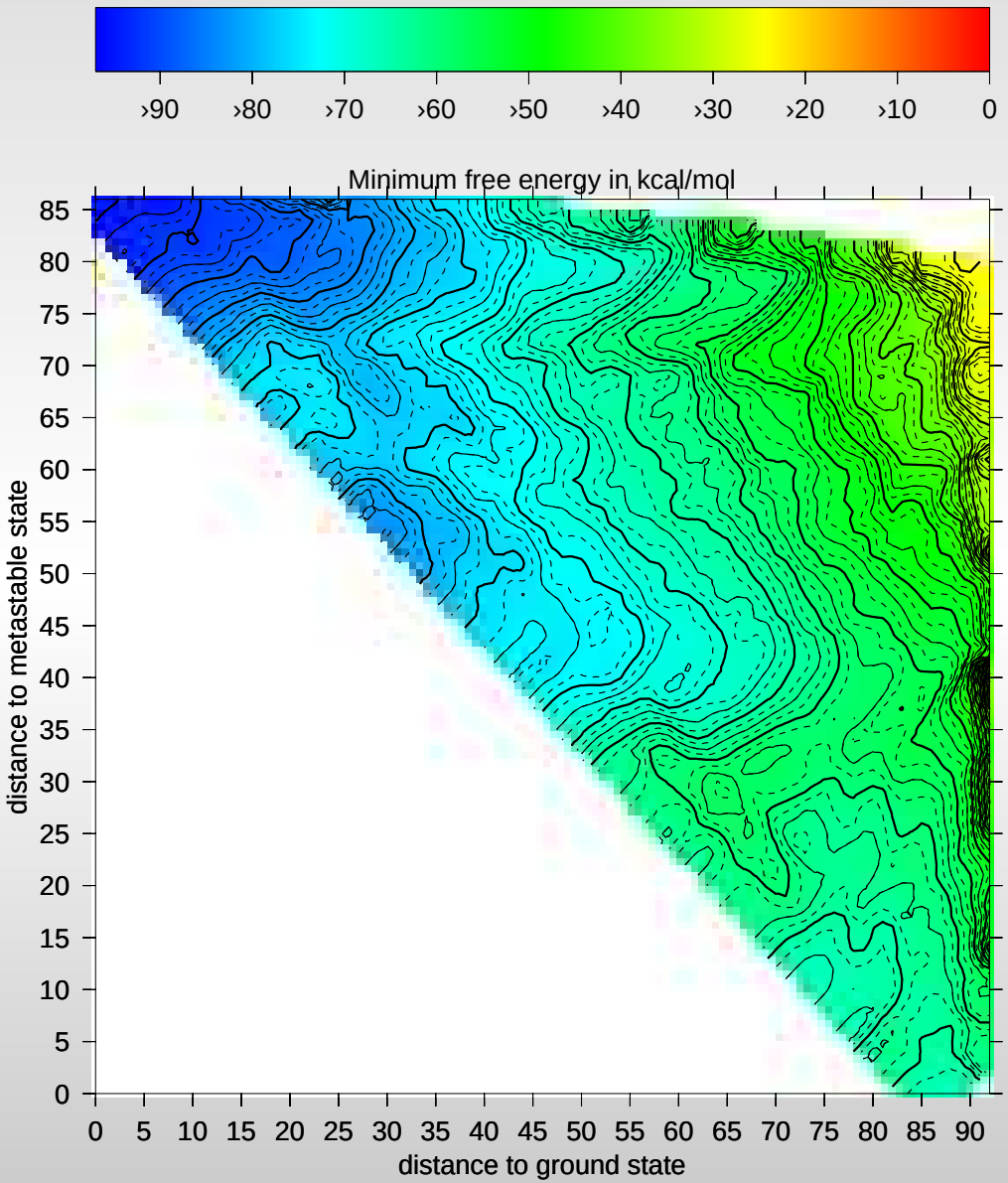
Example 2: SV11 RNA - distortion 10^5



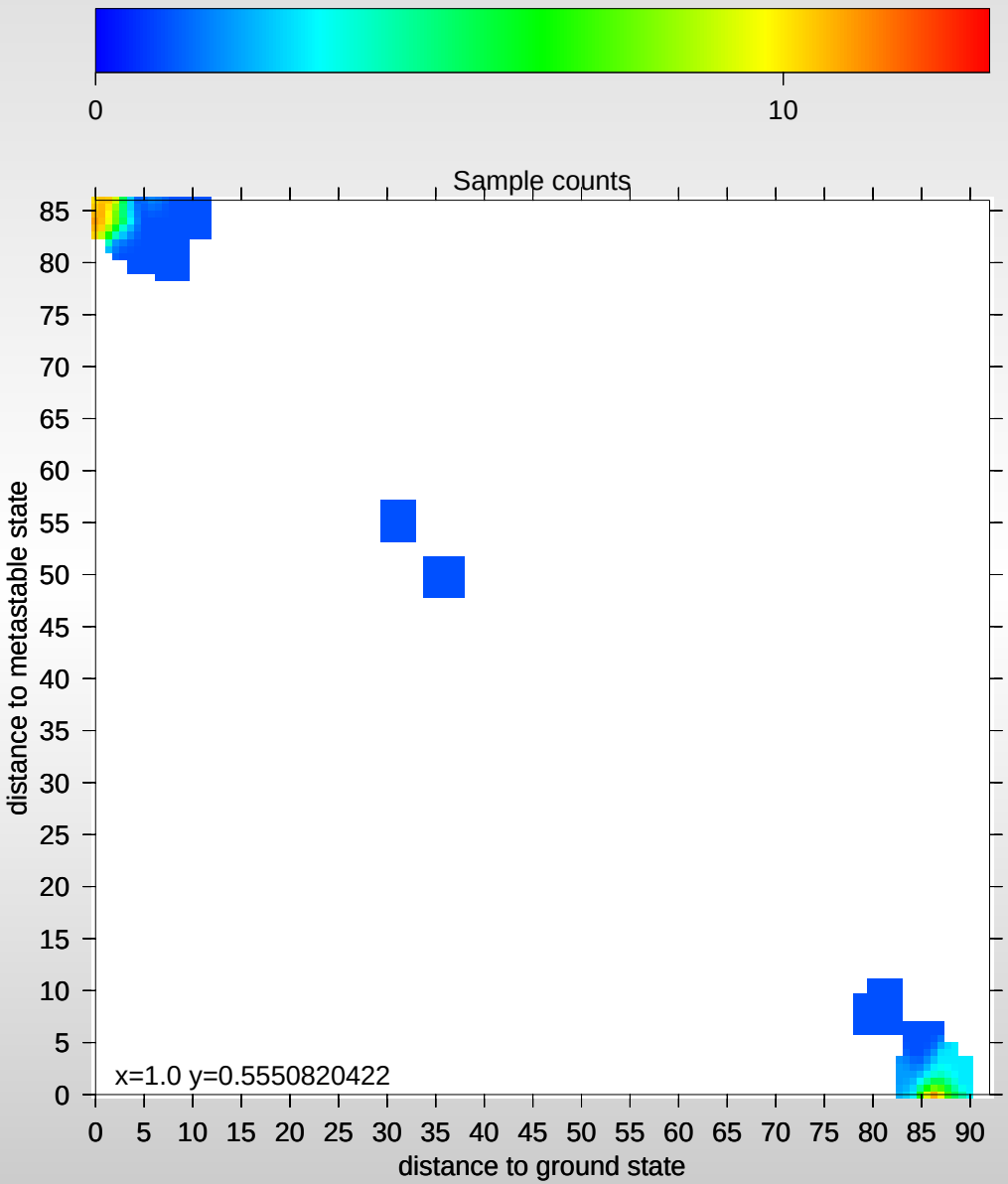
Example 2: SV11 RNA - distortion 10^6



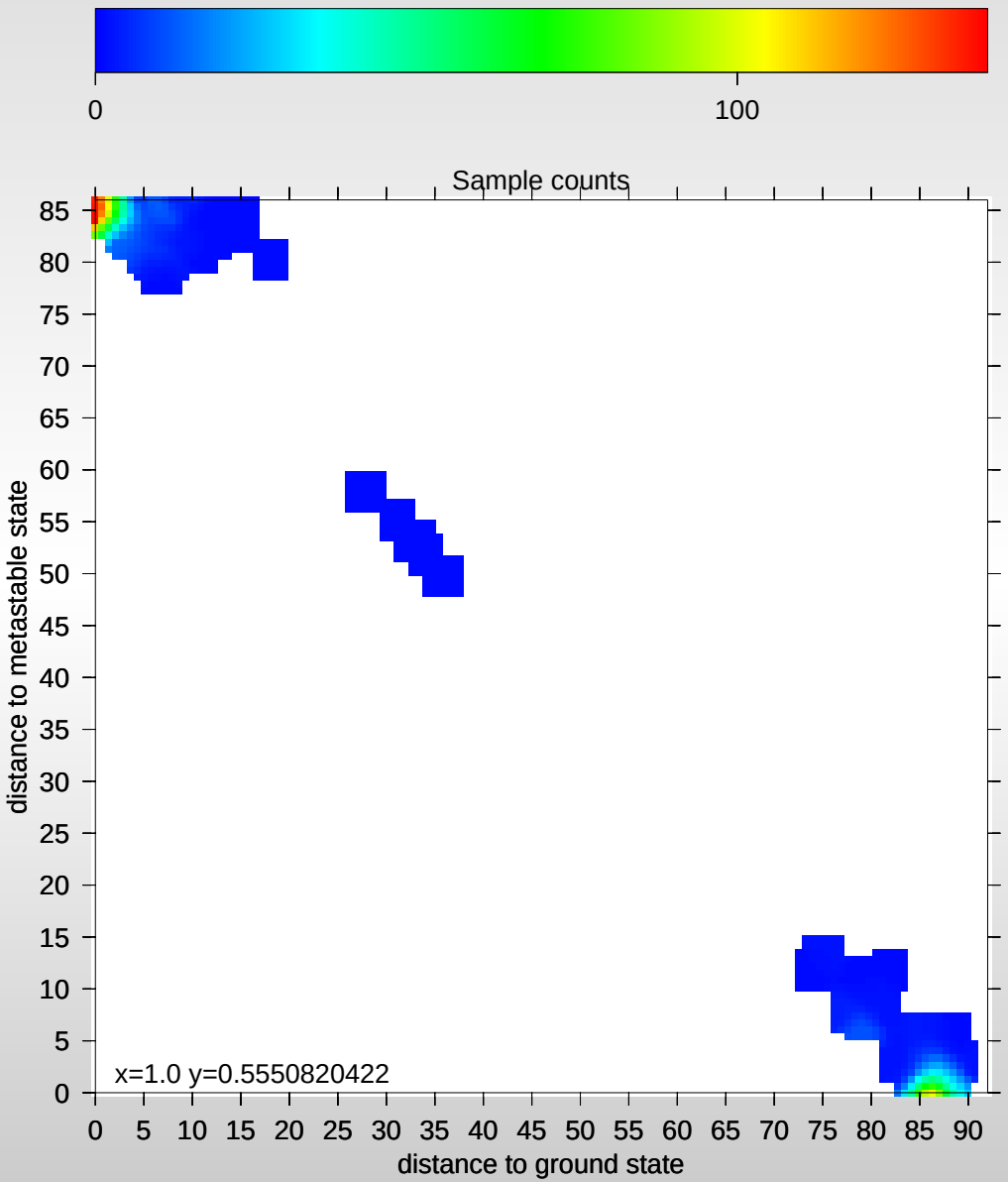
Example 2: SV11 RNA - RNA2Dfold



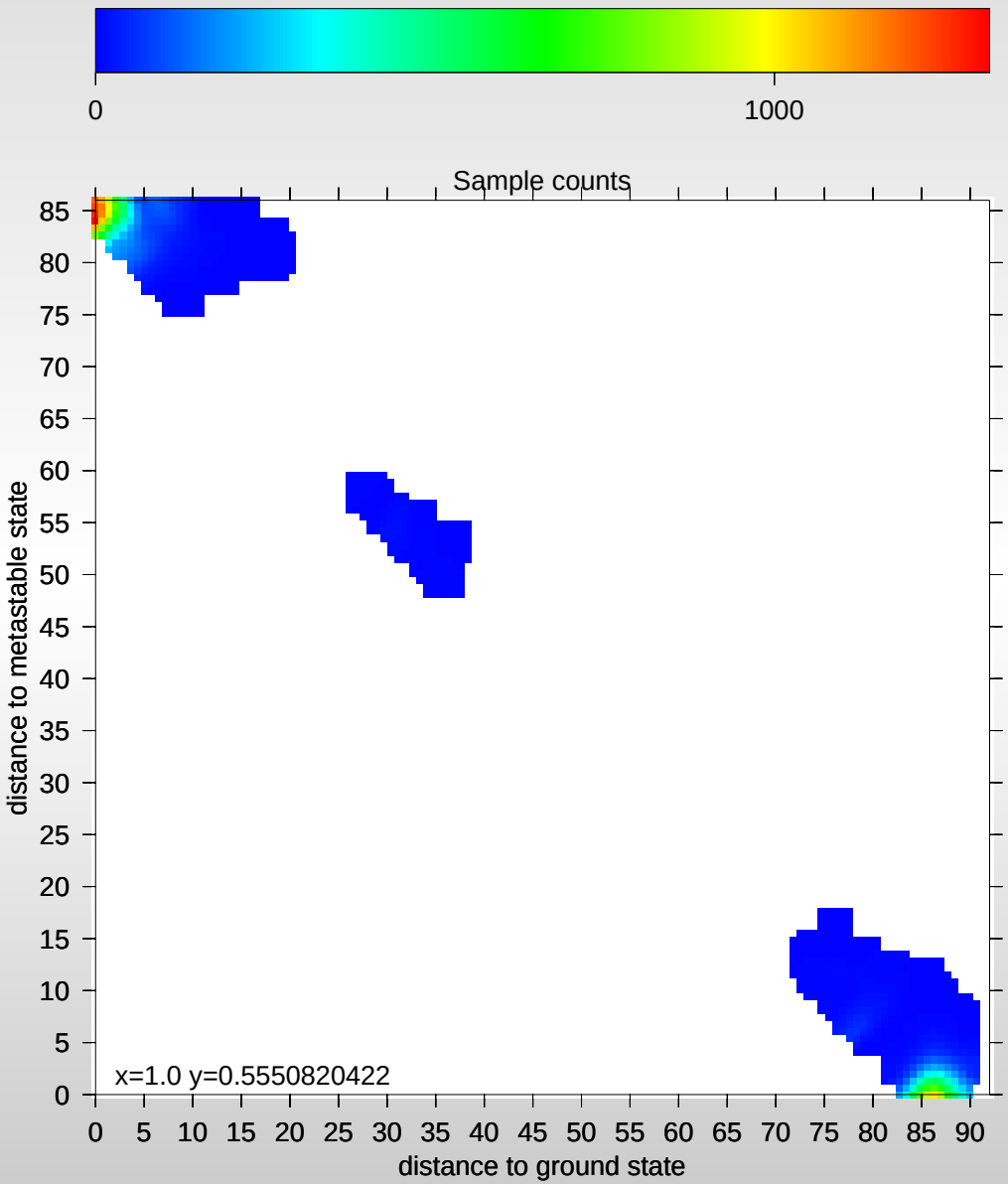
Example 2: SV11 RNA - distortion 10^2



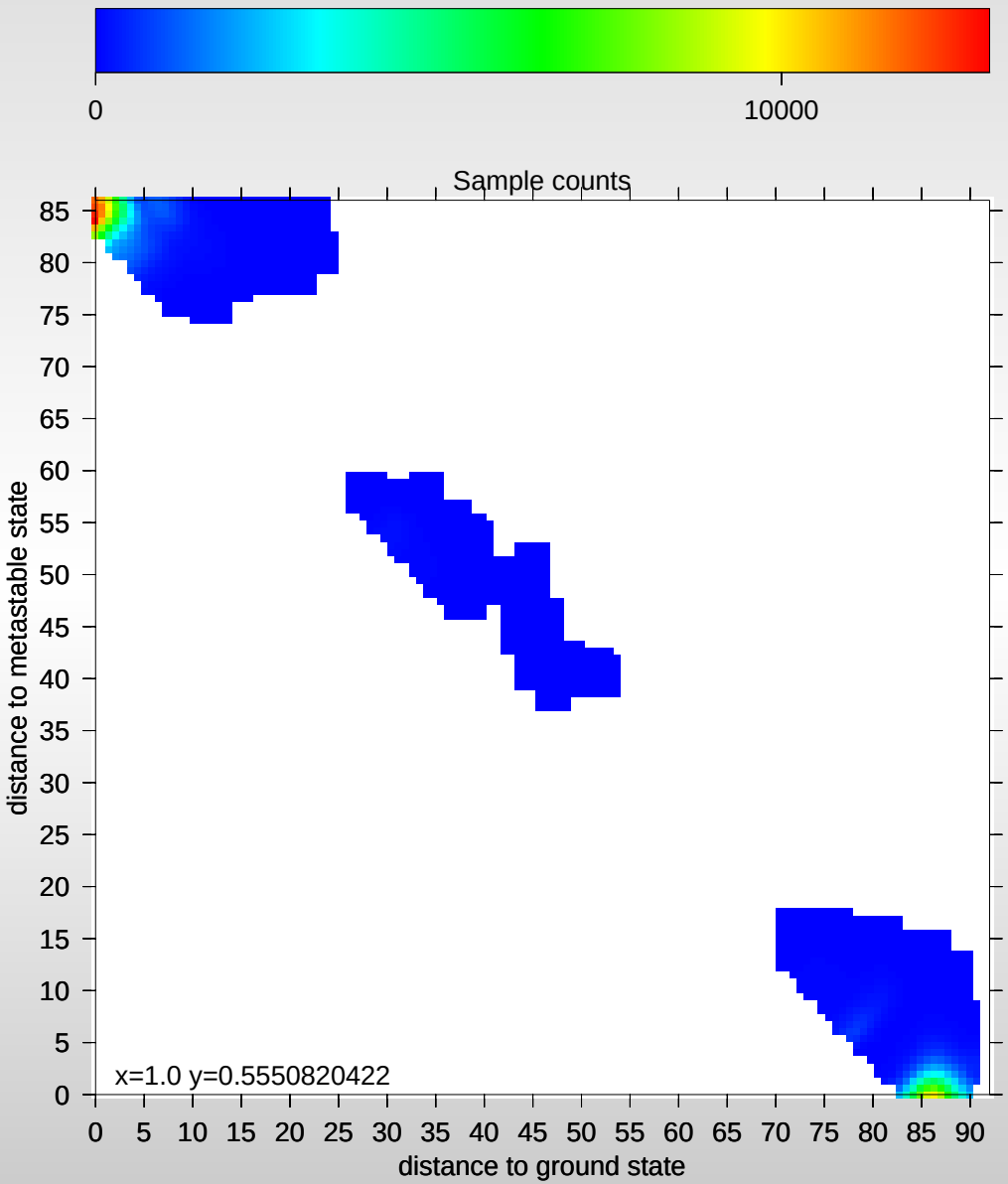
Example 2: SV11 RNA - distortion 10^3



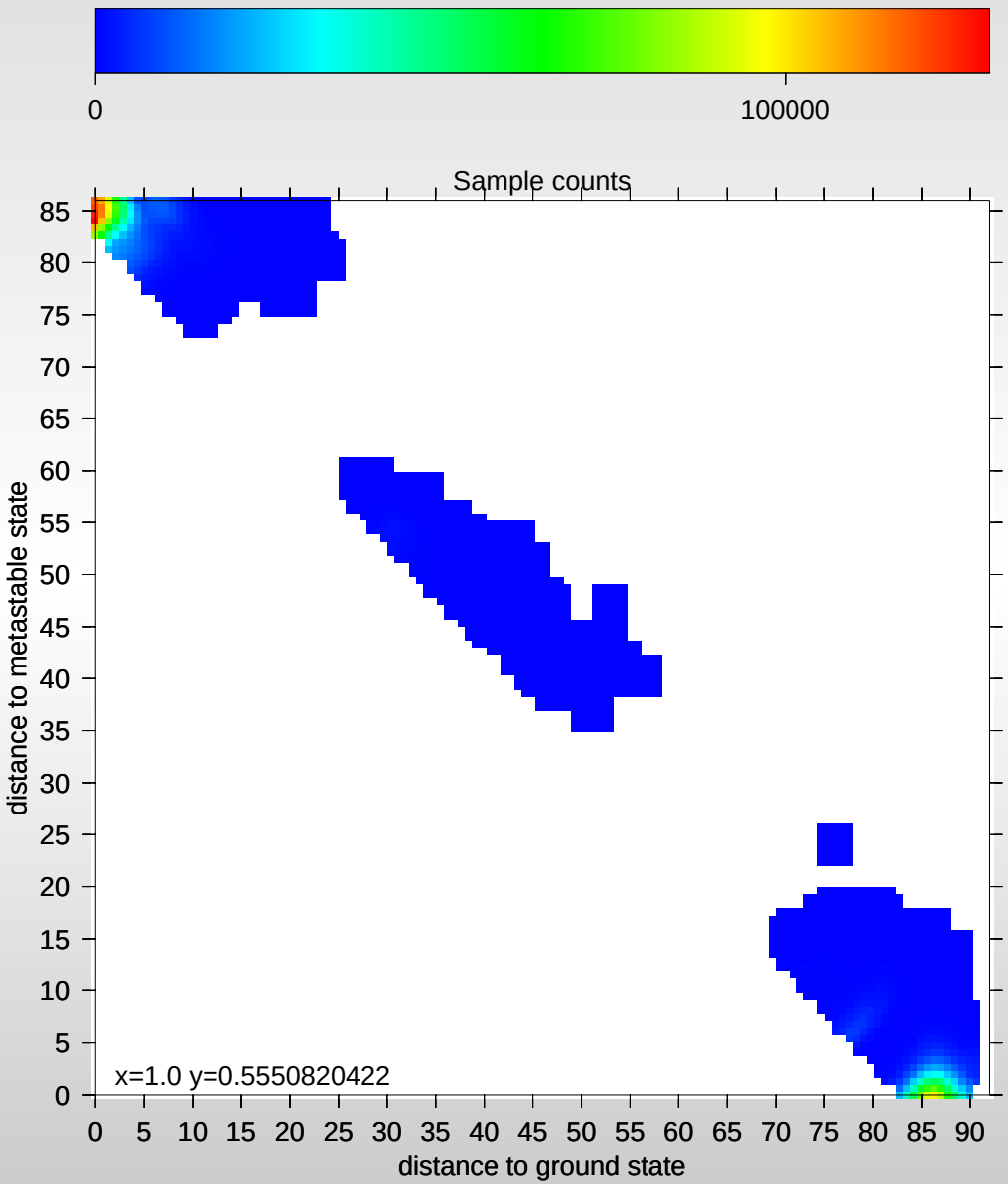
Example 2: SV11 RNA - distortion 10^4



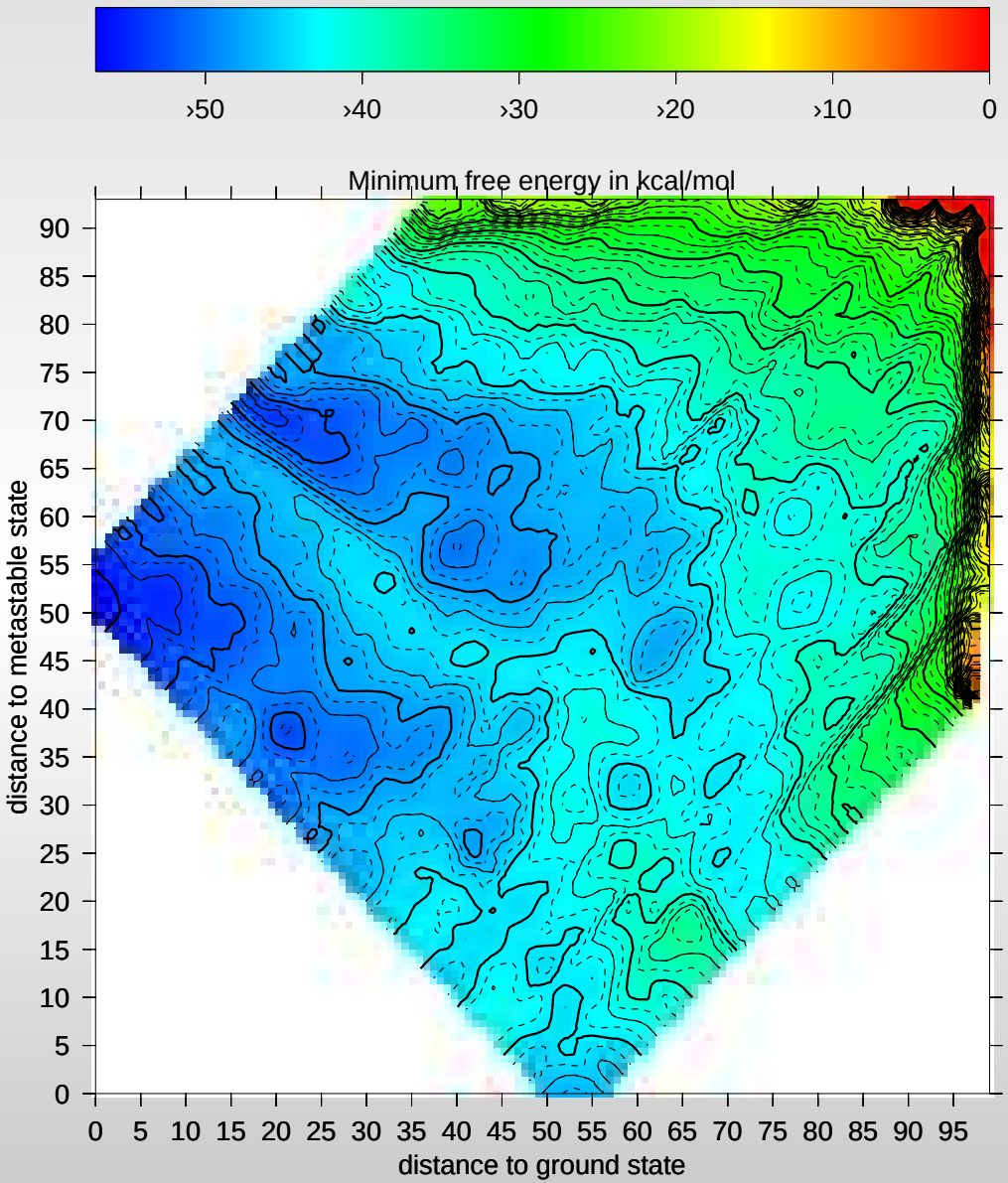
Example 2: SV11 RNA - distortion 10^5



Example 2: SV11 RNA - distortion 10^6

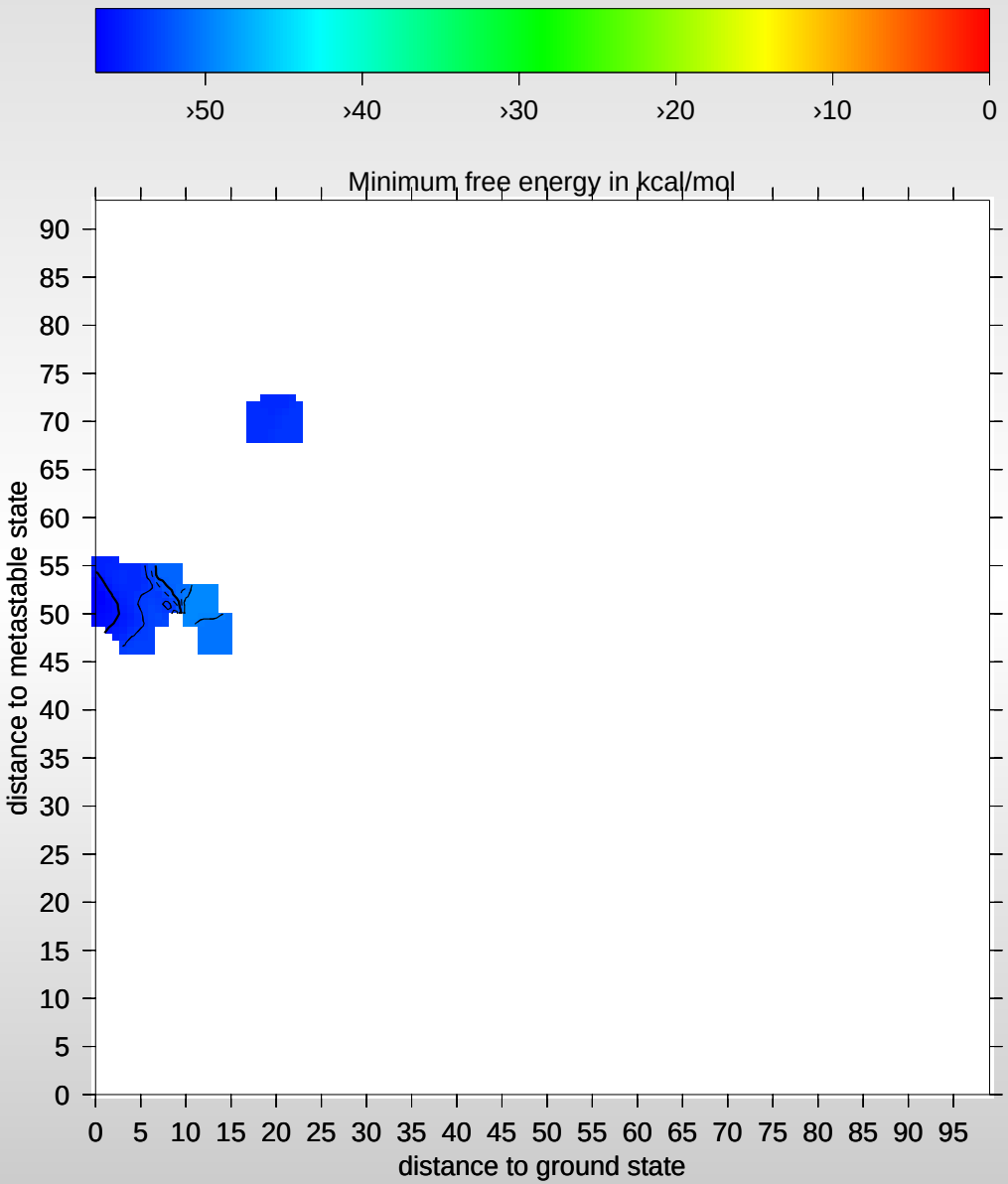


Example 3: 5'-UTR in MS2[11] - RNA2Dfold

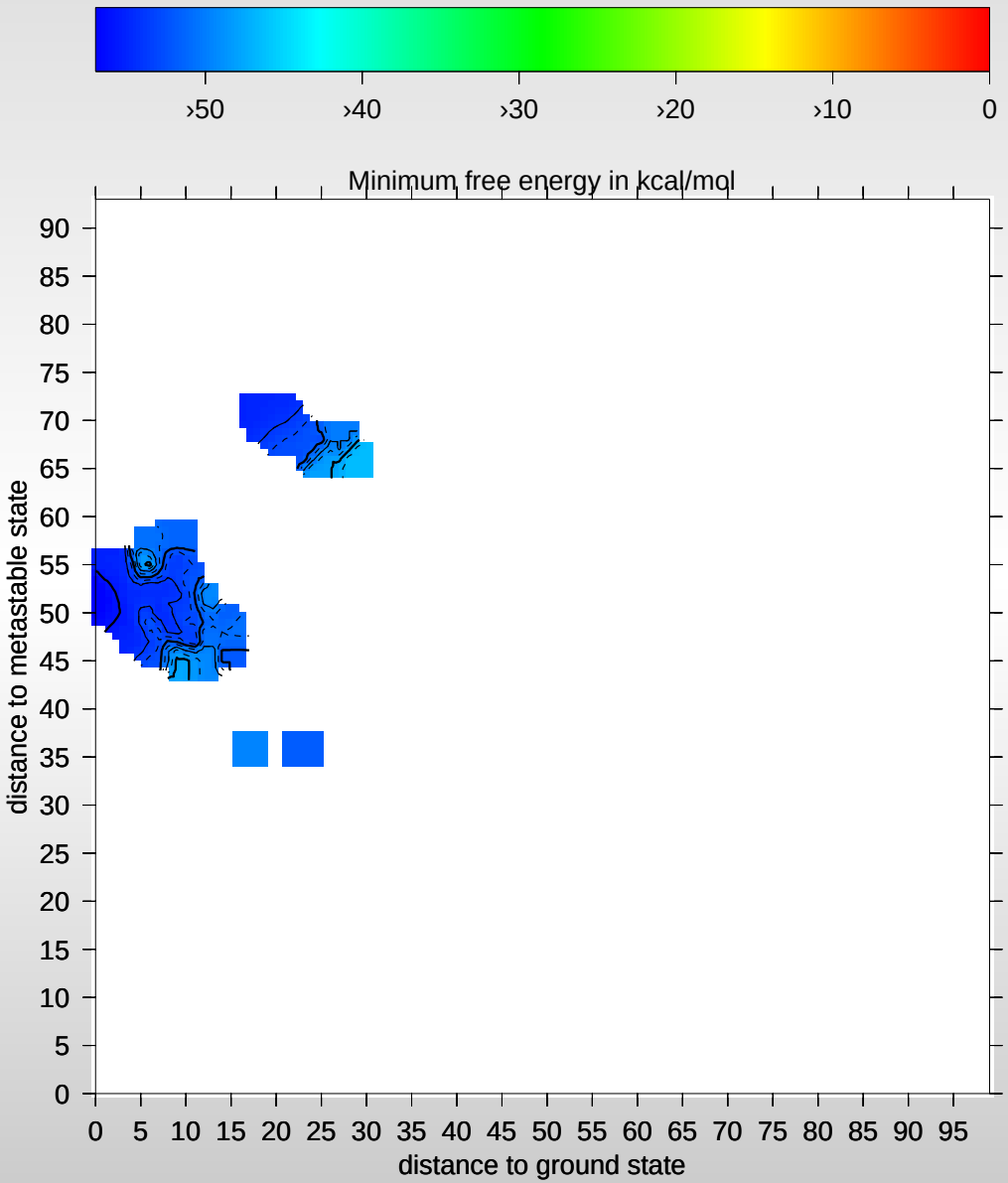


[11] van Meerten et al. 2001

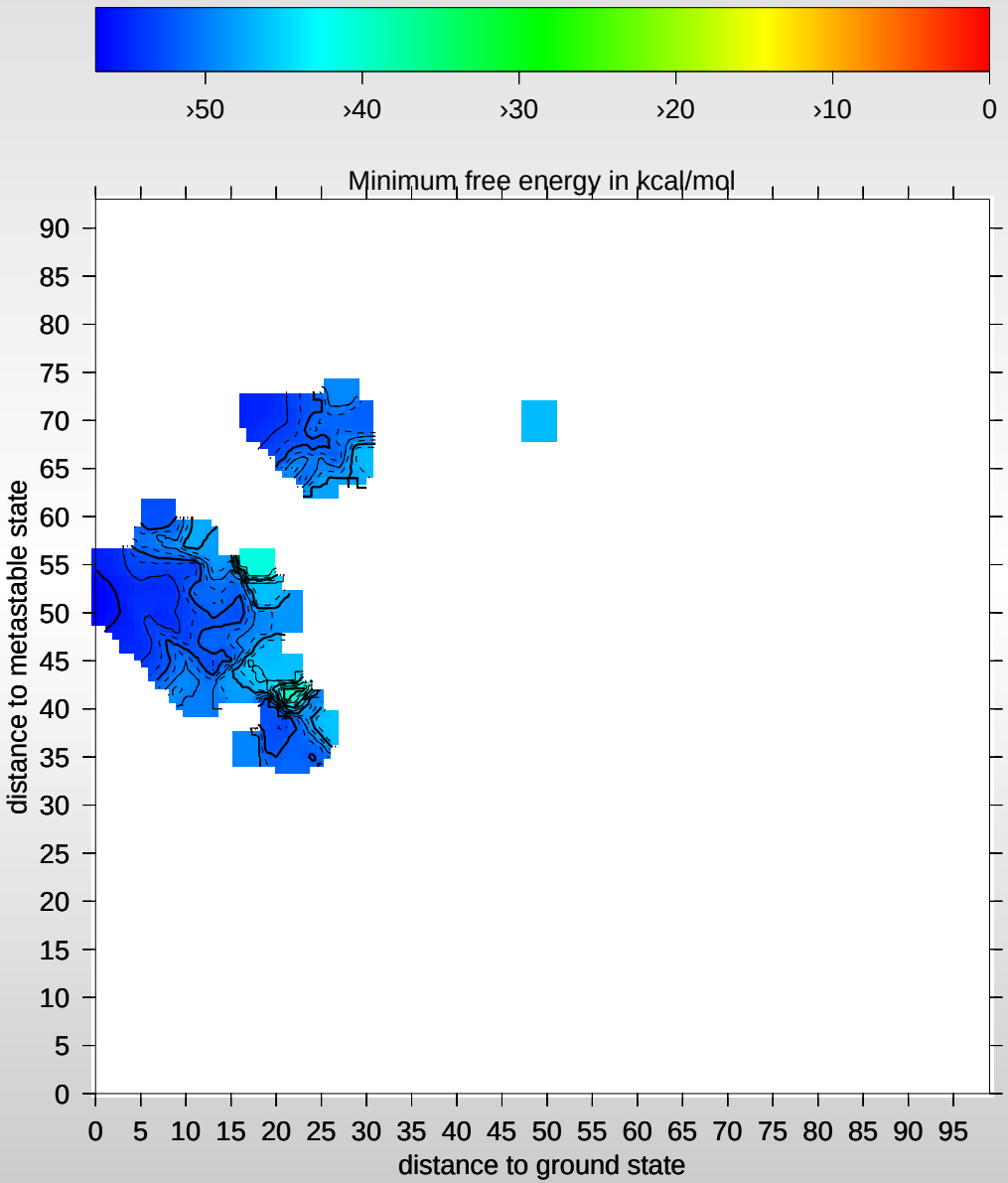
Example 3: 5'-UTR in MS2 - RNAsubopt -p 10^2



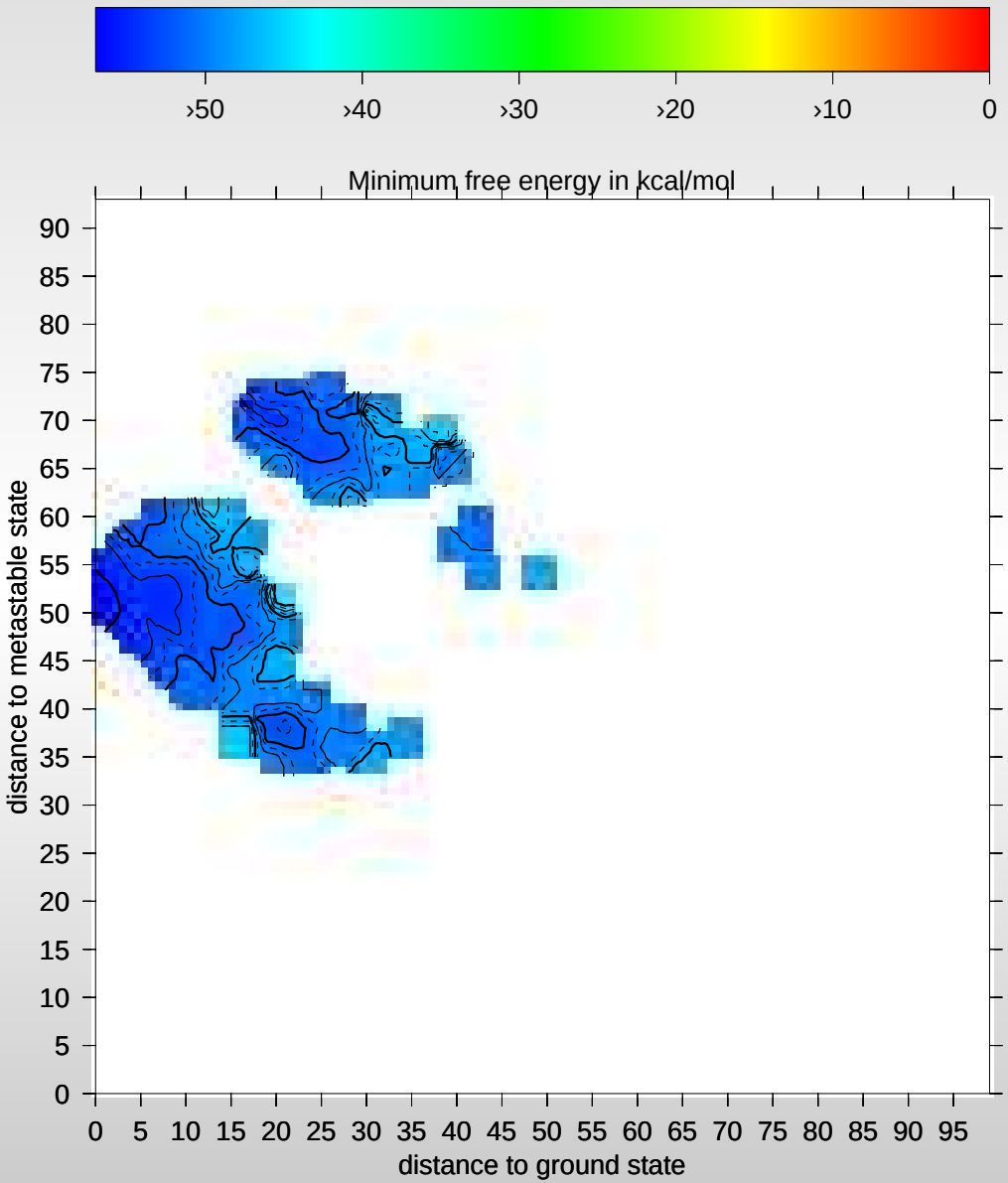
Example 3: 5’-UTR in MS2 - RNAsubopt -p 10^3



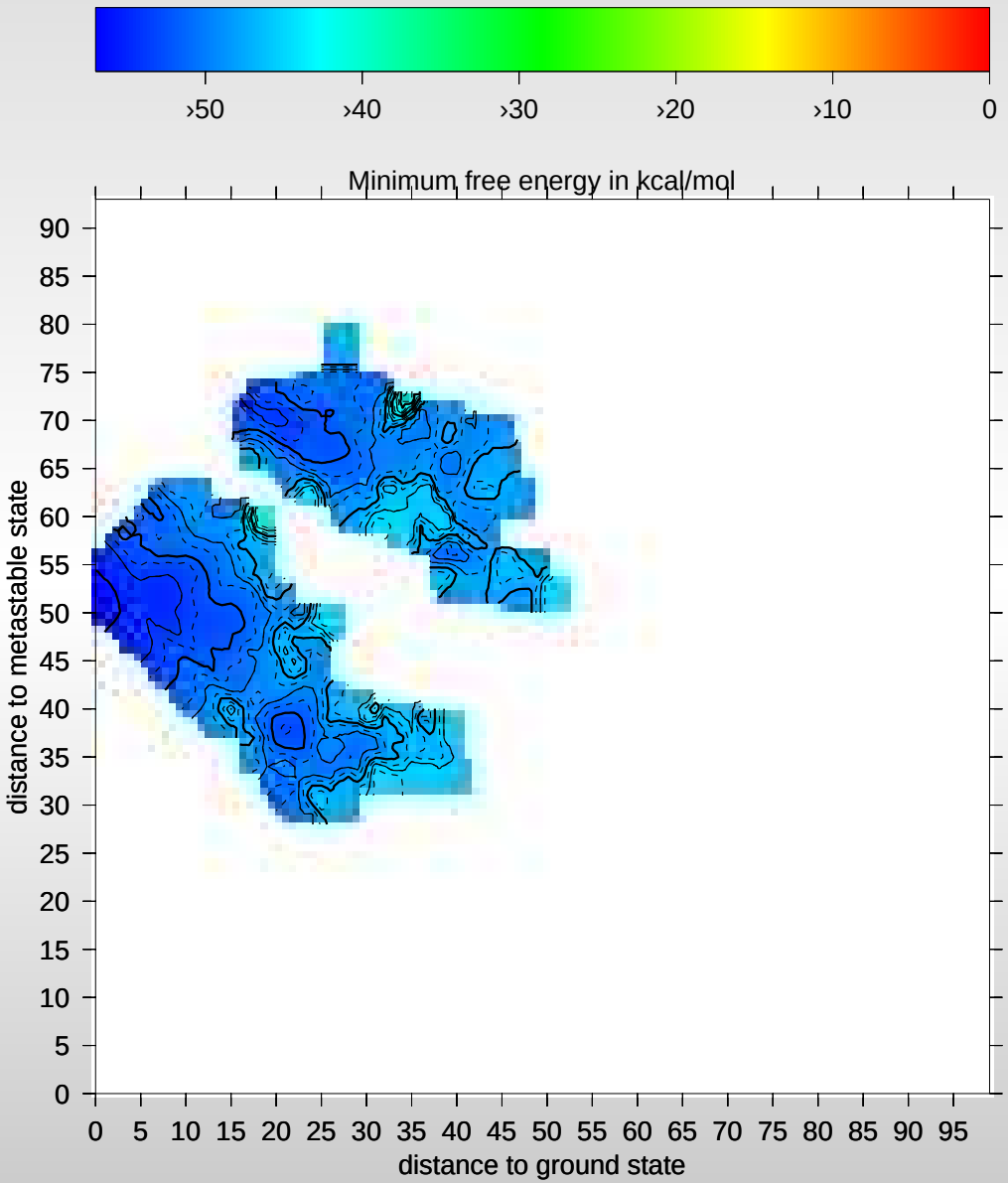
Example 3: 5'-UTR in MS2 - RNAsubopt -p 10^4



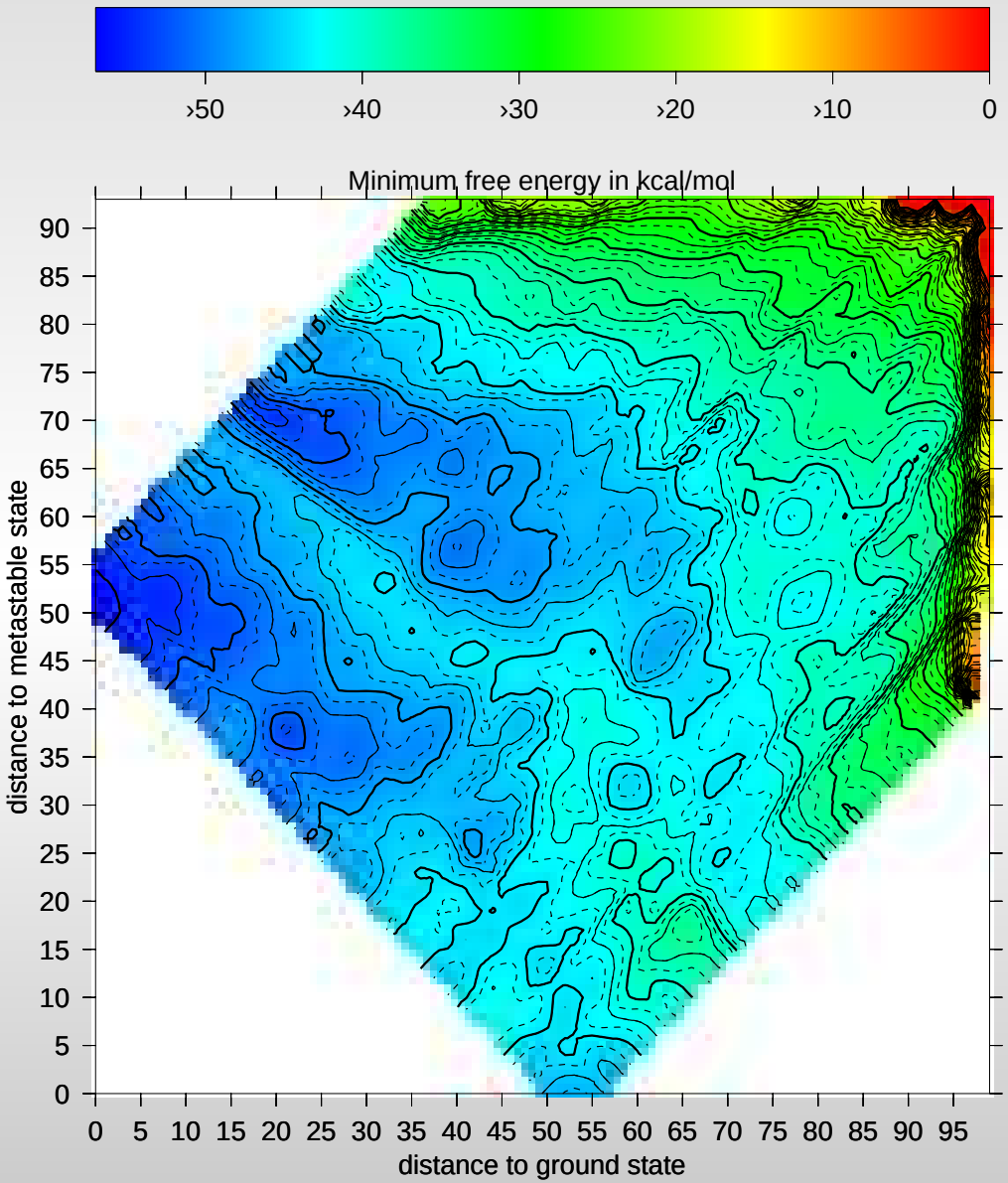
Example 3: 5'-UTR in MS2 - RNAsubopt -p 10^5



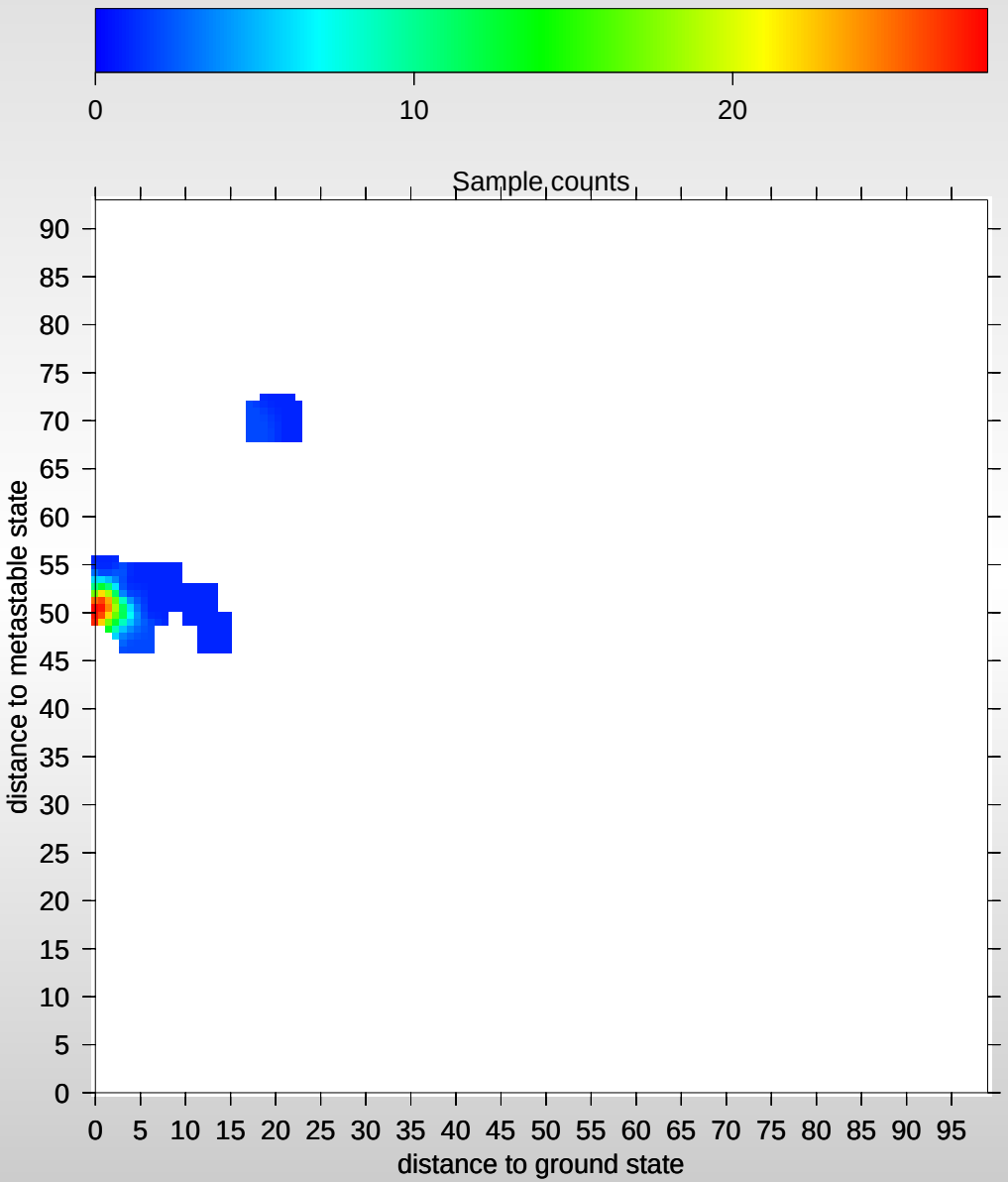
Example 3: 5'-UTR in MS2 - RNAsubopt -p 10^6



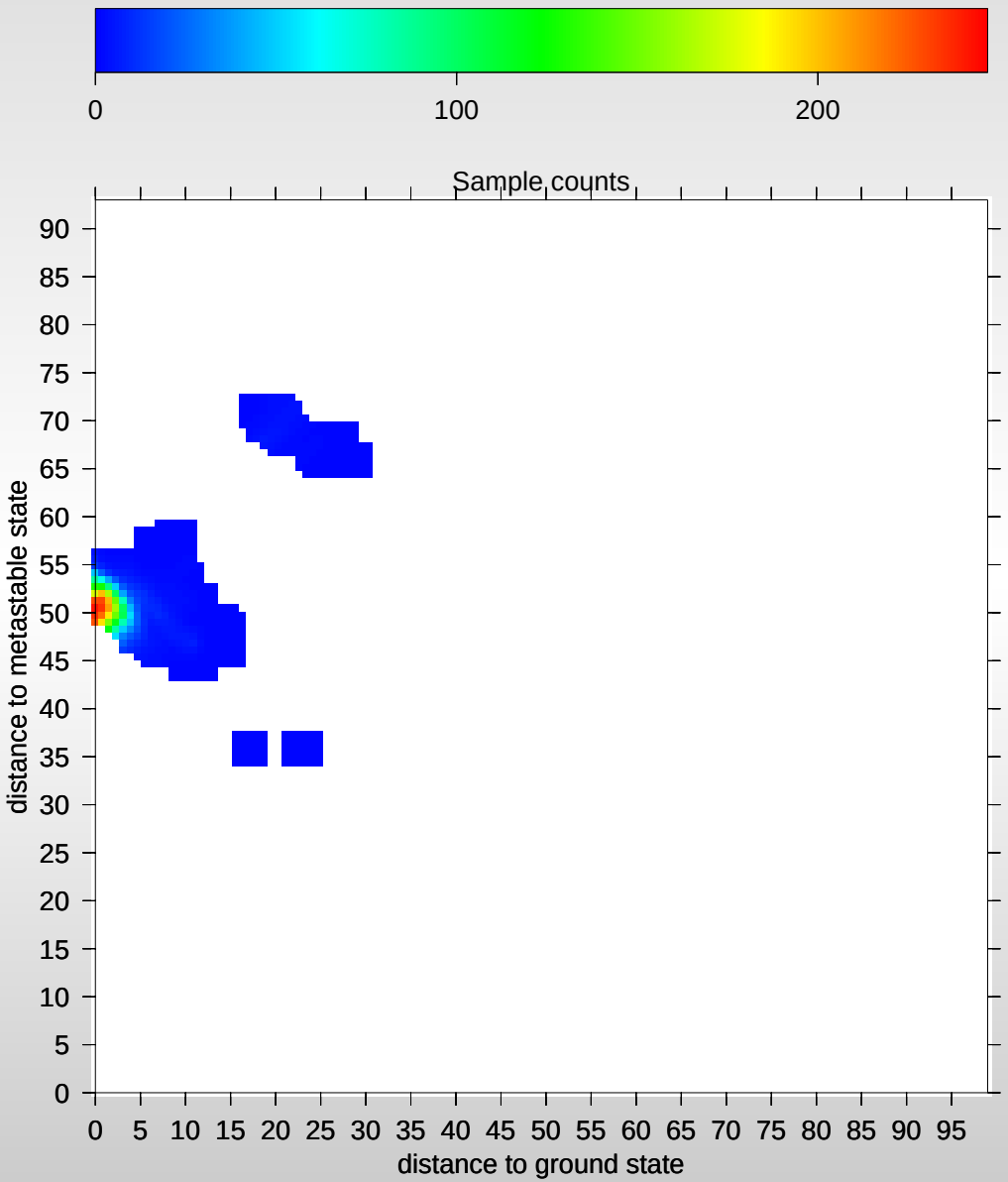
Example 3: 5’-UTR in MS2 - RNA2Dfold



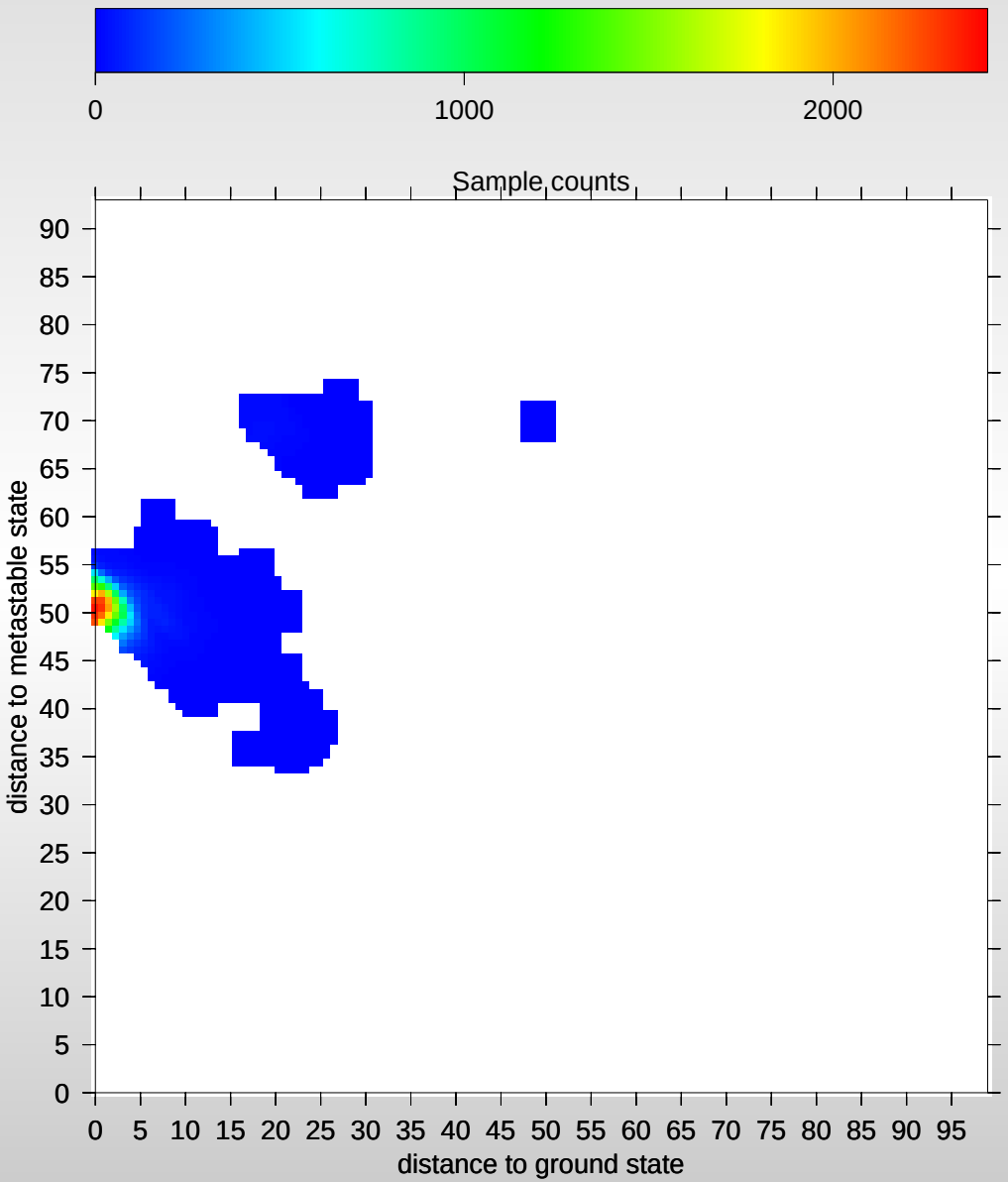
Example 3: 5'-UTR in MS2 - RNAsubopt -p 10^2



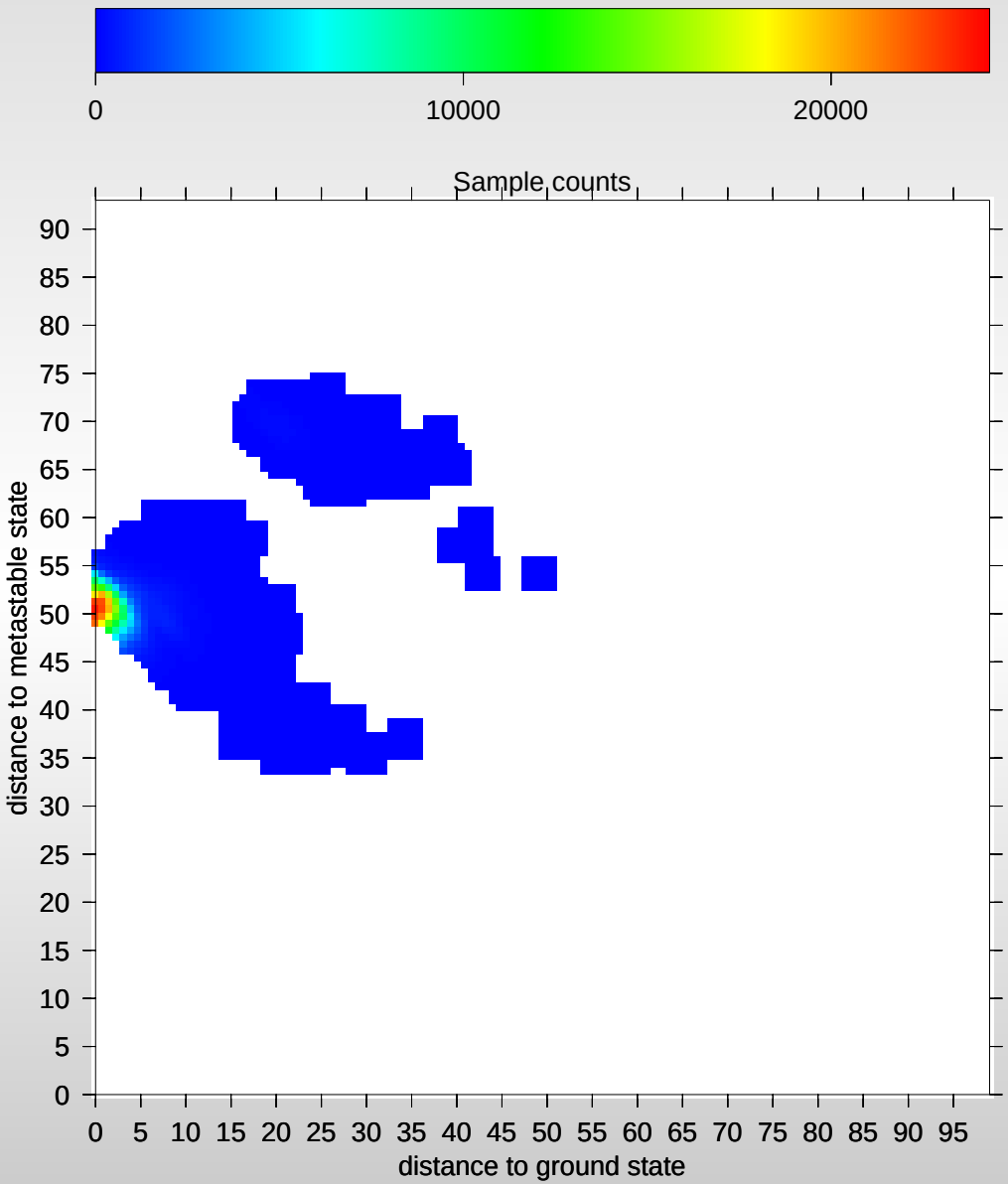
Example 3: 5’-UTR in MS2 - RNAsubopt -p 10^3



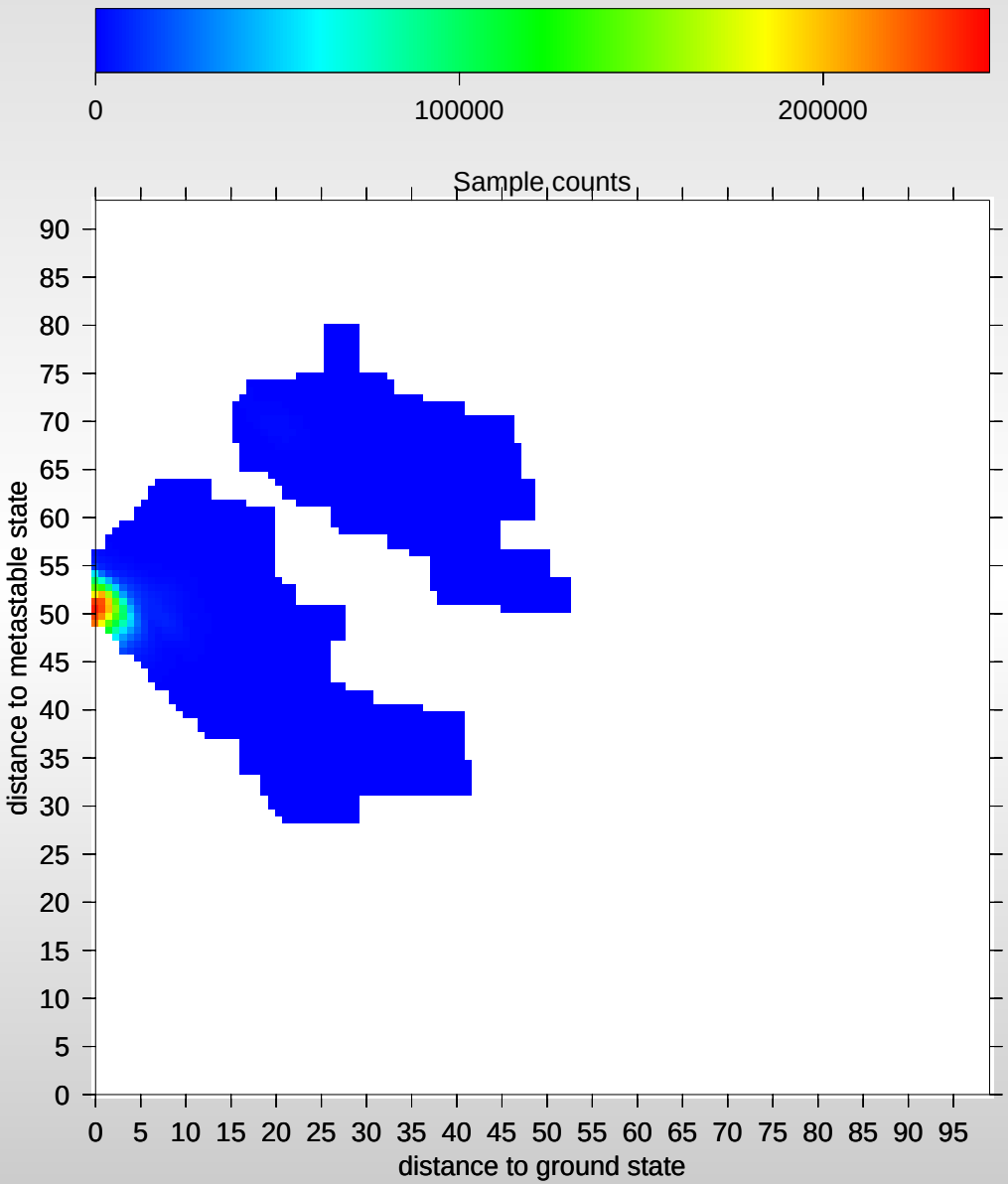
Example 3: 5'-UTR in MS2 - RNAsubopt -p 10^4



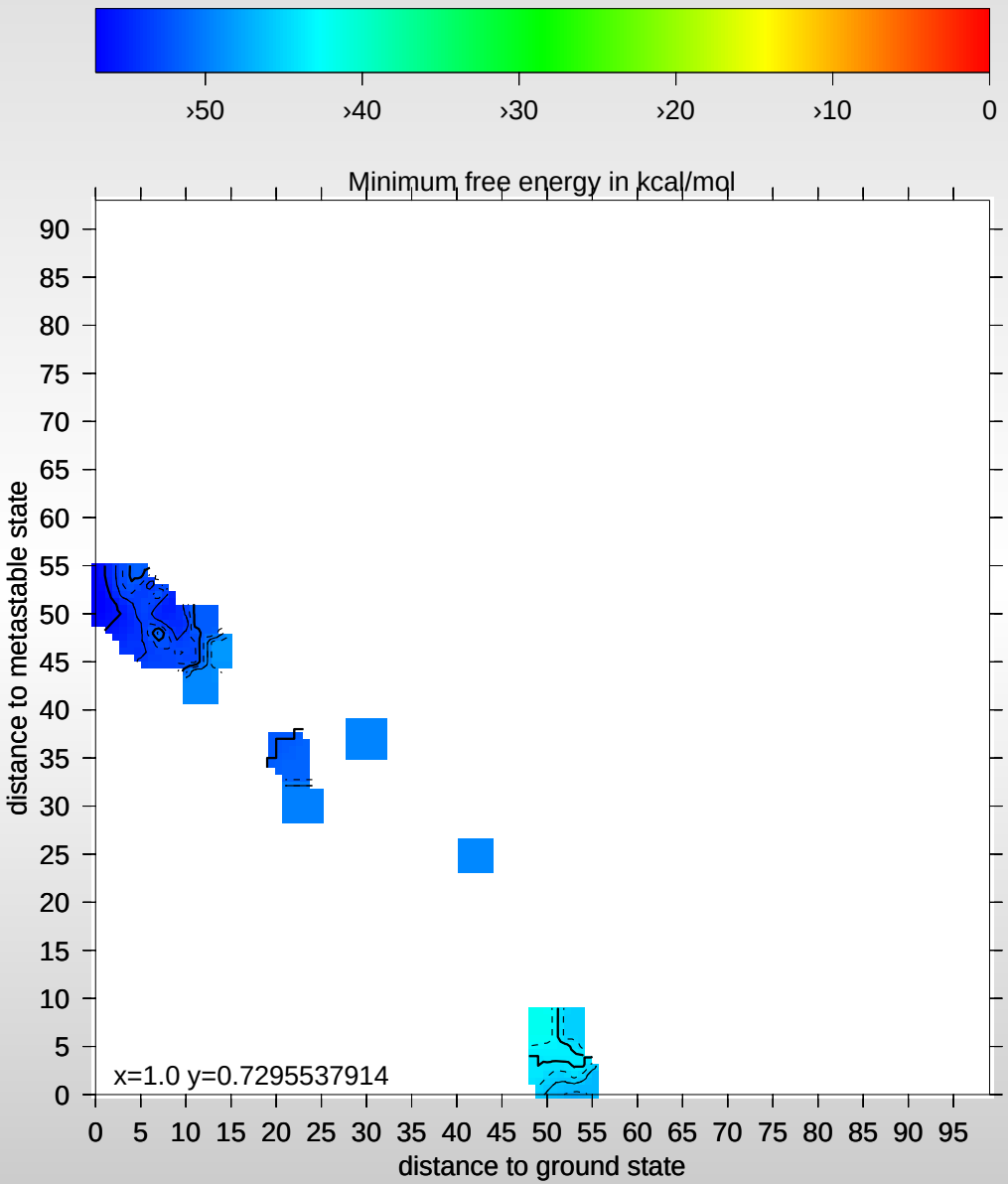
Example 3: 5’-UTR in MS2 - RNAsubopt -p 10^5



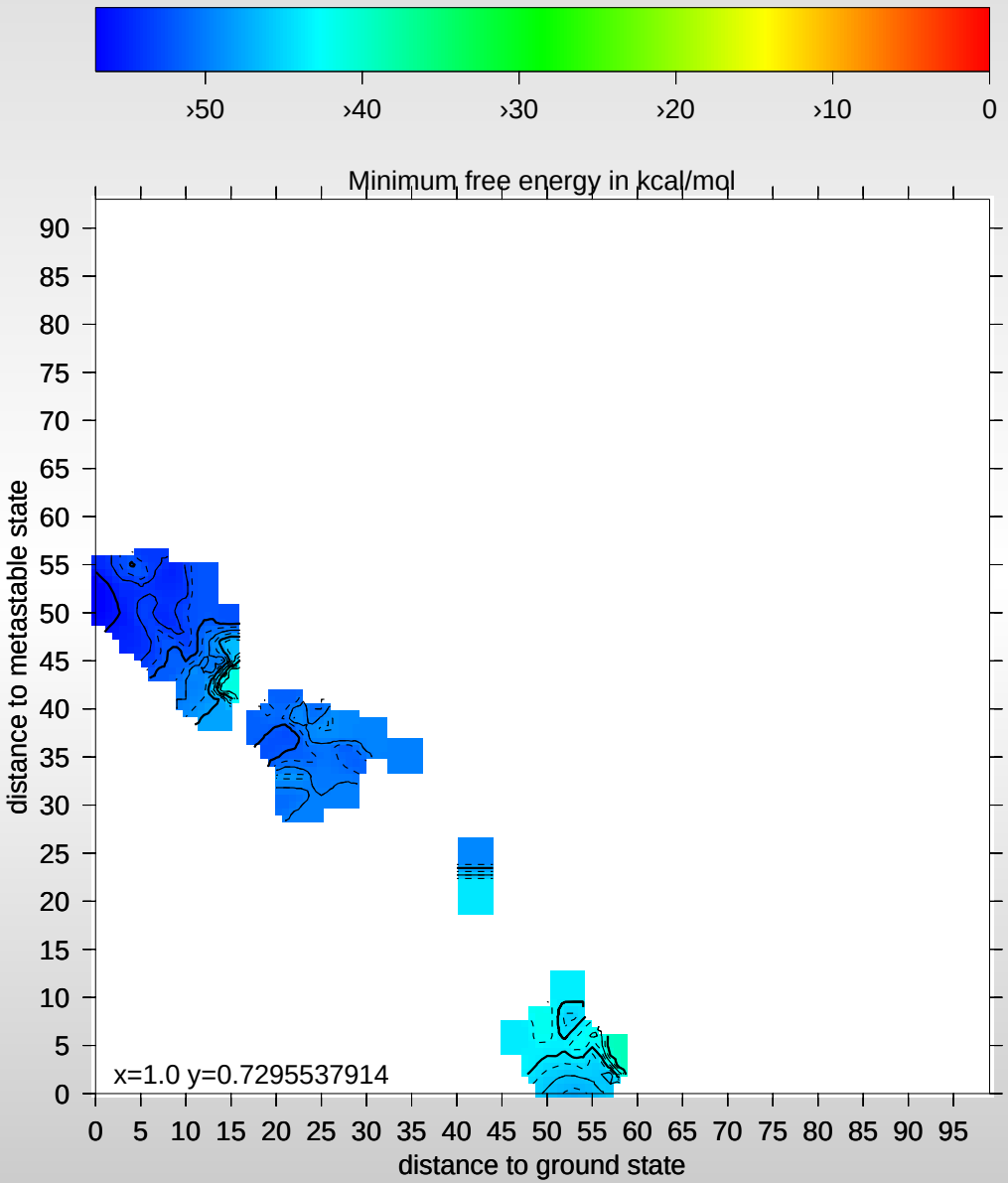
Example 3: 5'-UTR in MS2 - RNAsubopt -p 10^6



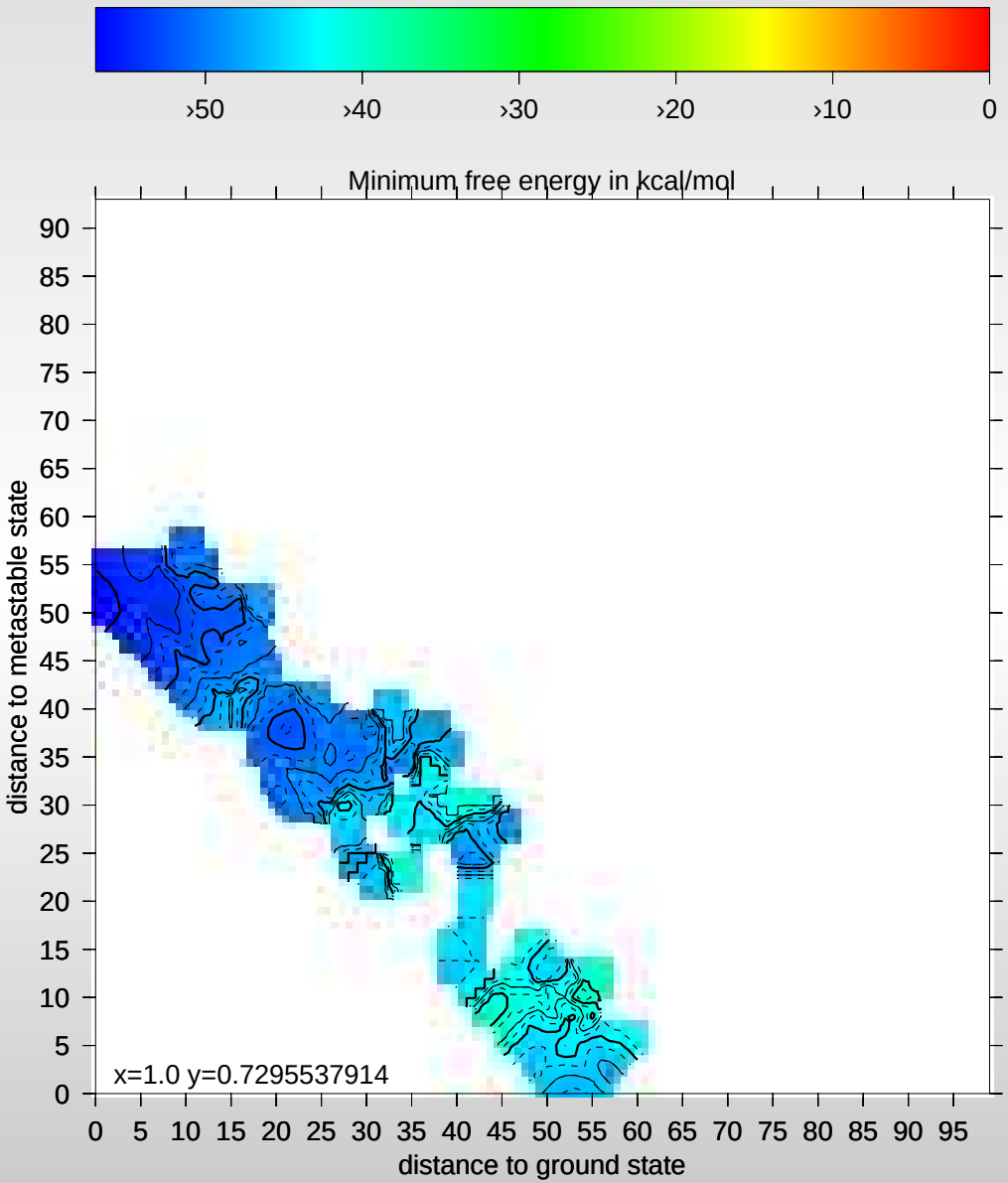
Example 3: 5'-UTR in MS2 - distortion 10^2



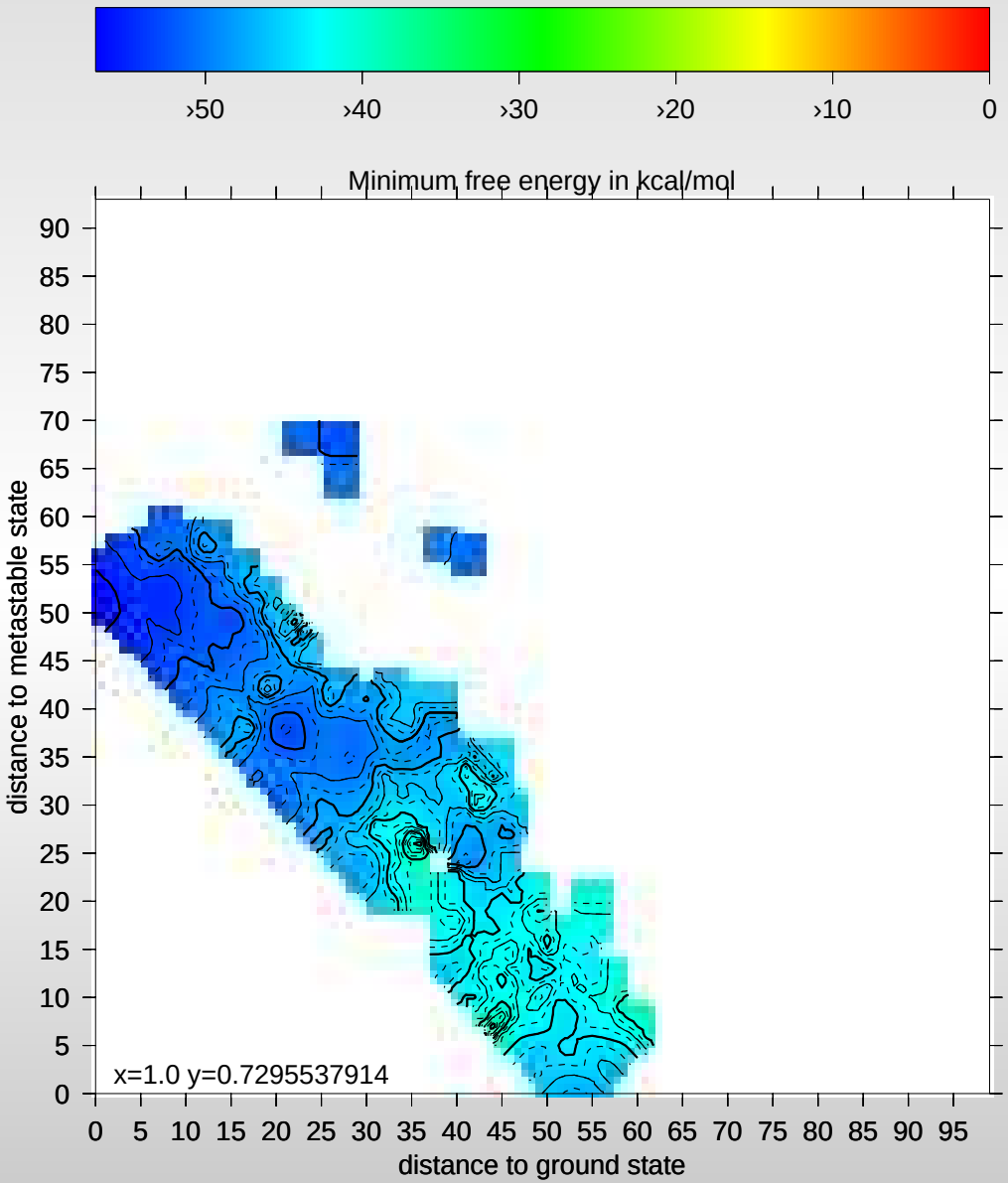
Example 3: 5'-UTR in MS2 - distortion 10^3



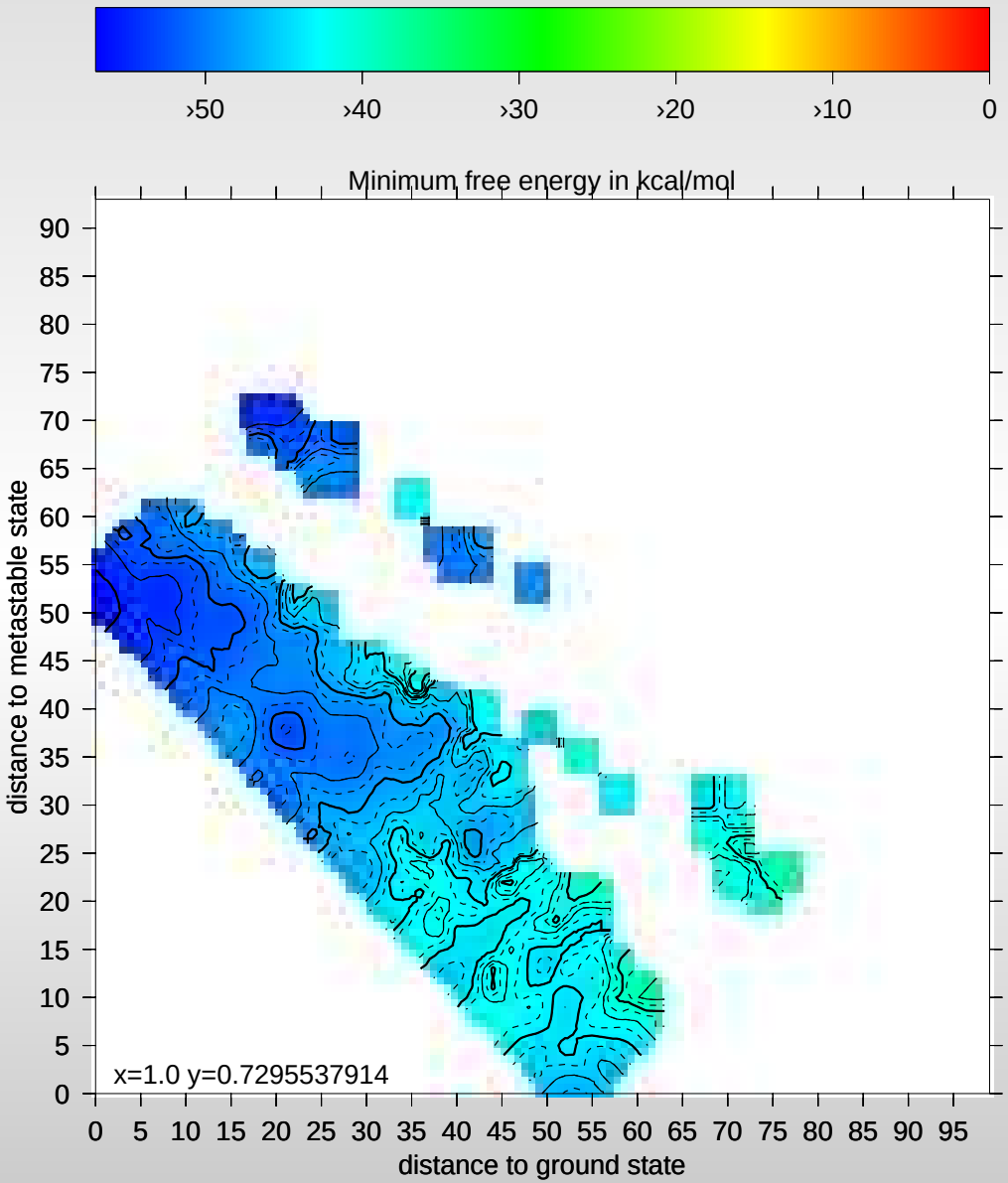
Example 3: 5'-UTR in MS2 - distortion 10^4



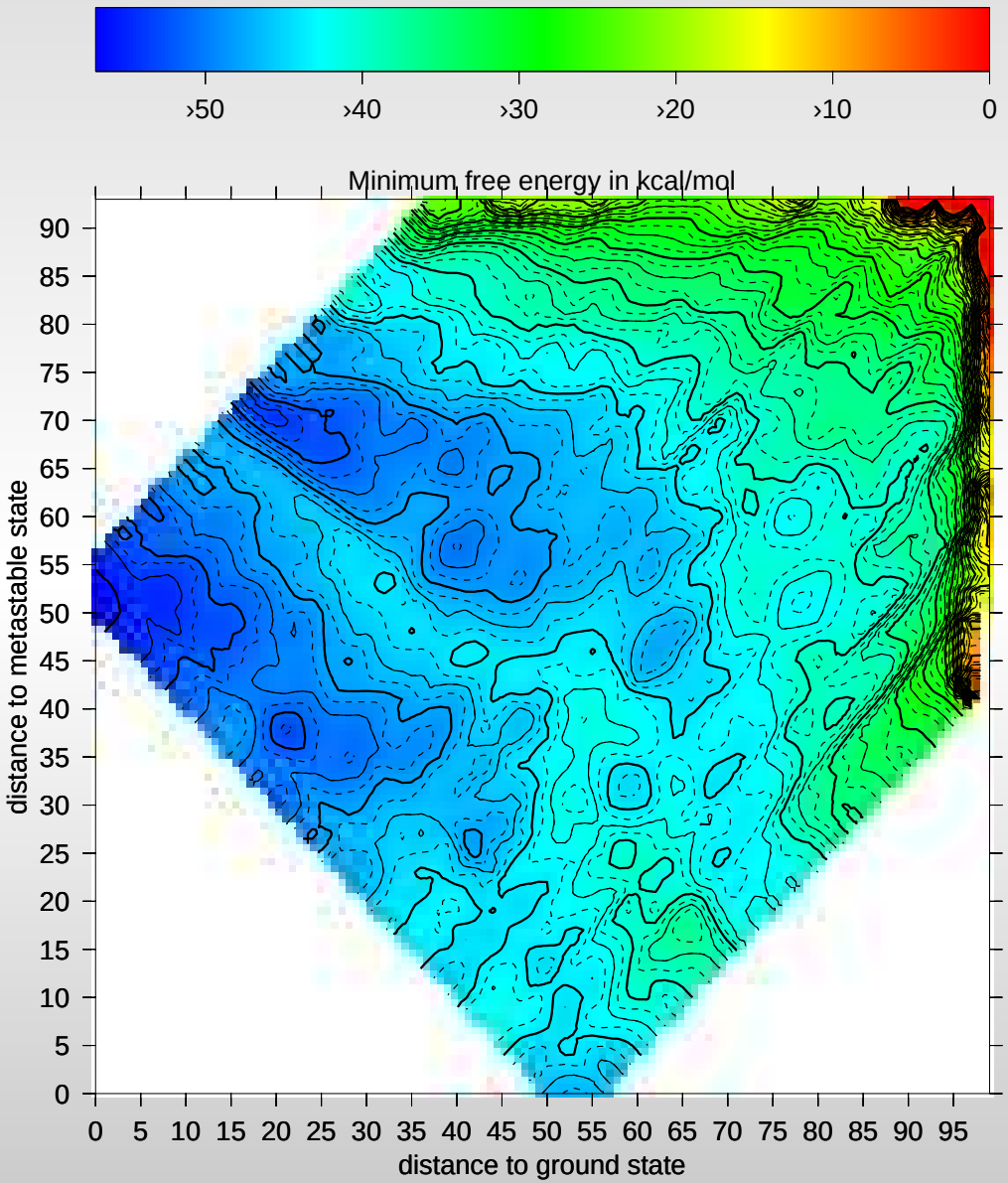
Example 3: 5'-UTR in MS2 - distortion 10^5



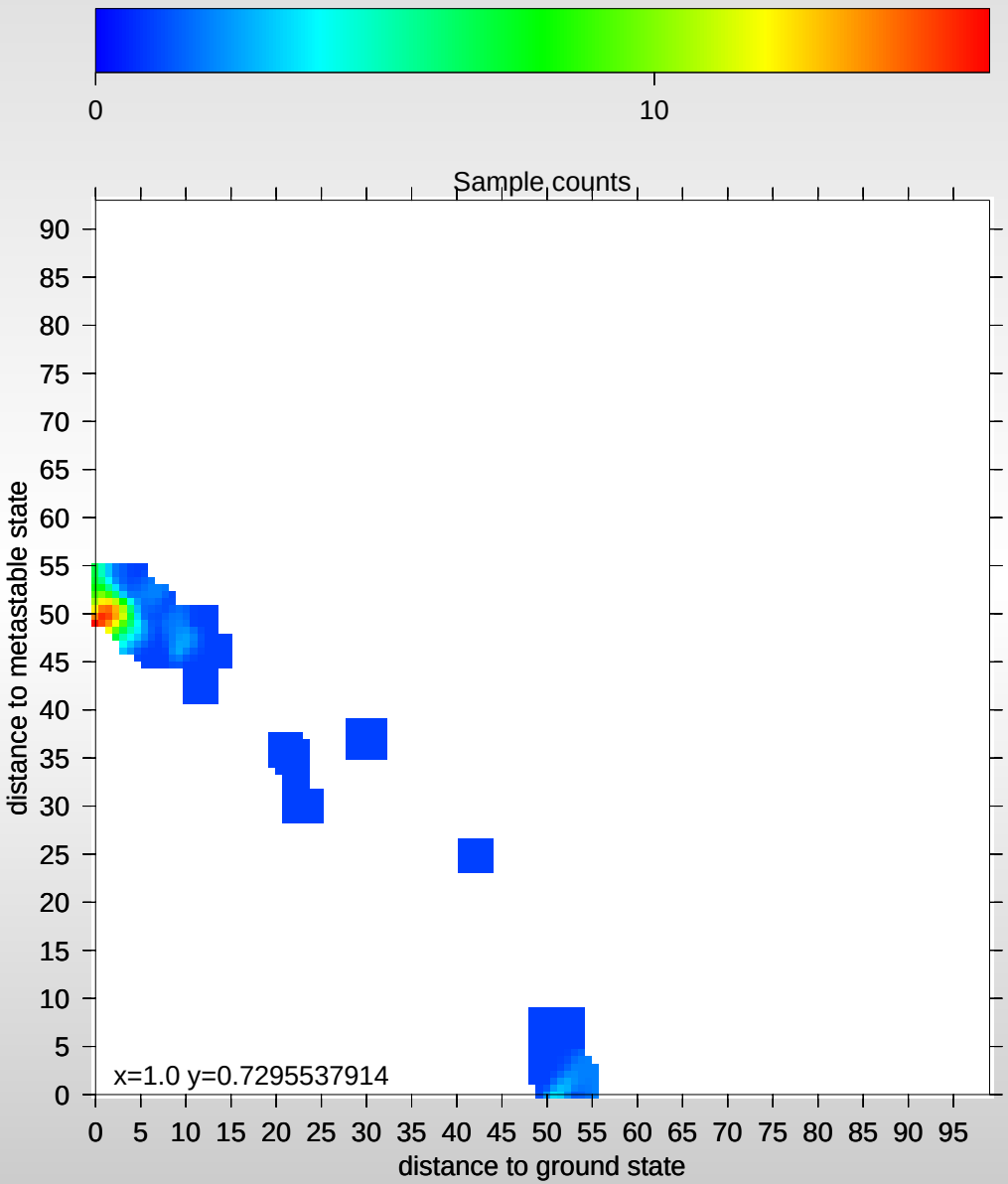
Example 3: 5'-UTR in MS2 - distortion 10^6



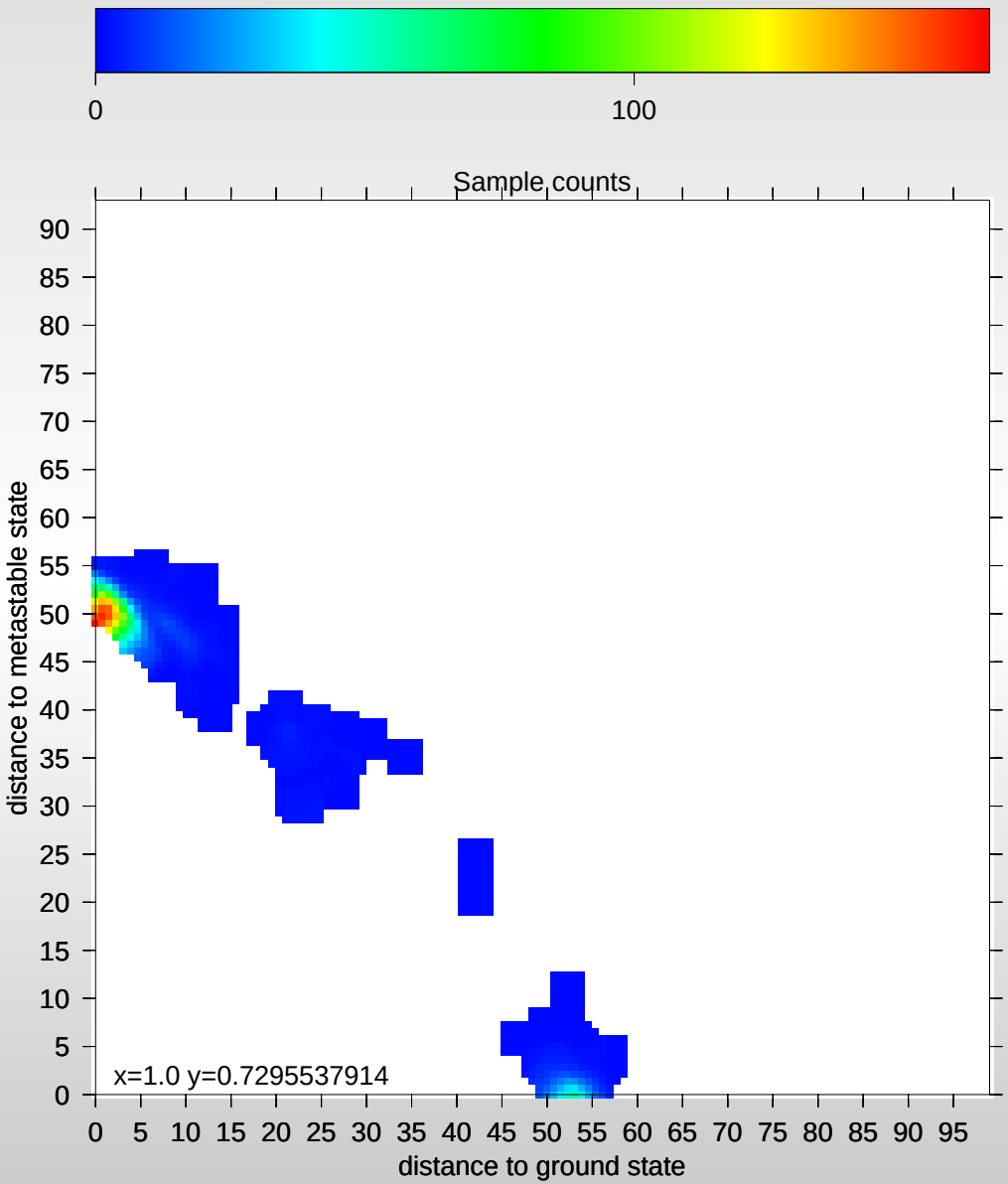
Example 3: 5’-UTR in MS2 - RNA2Dfold



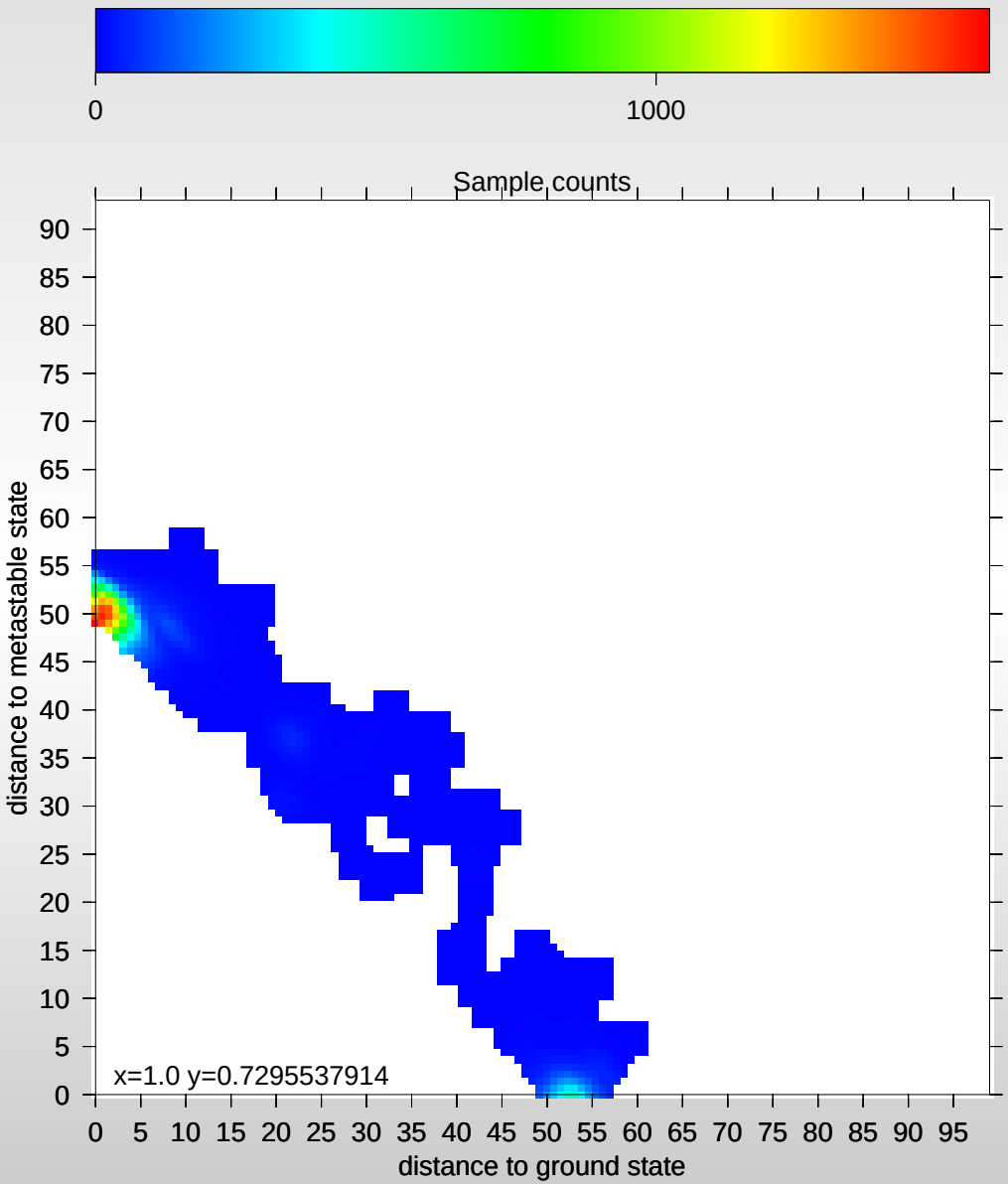
Example 3: 5’-UTR in MS2 - distortion 10^2



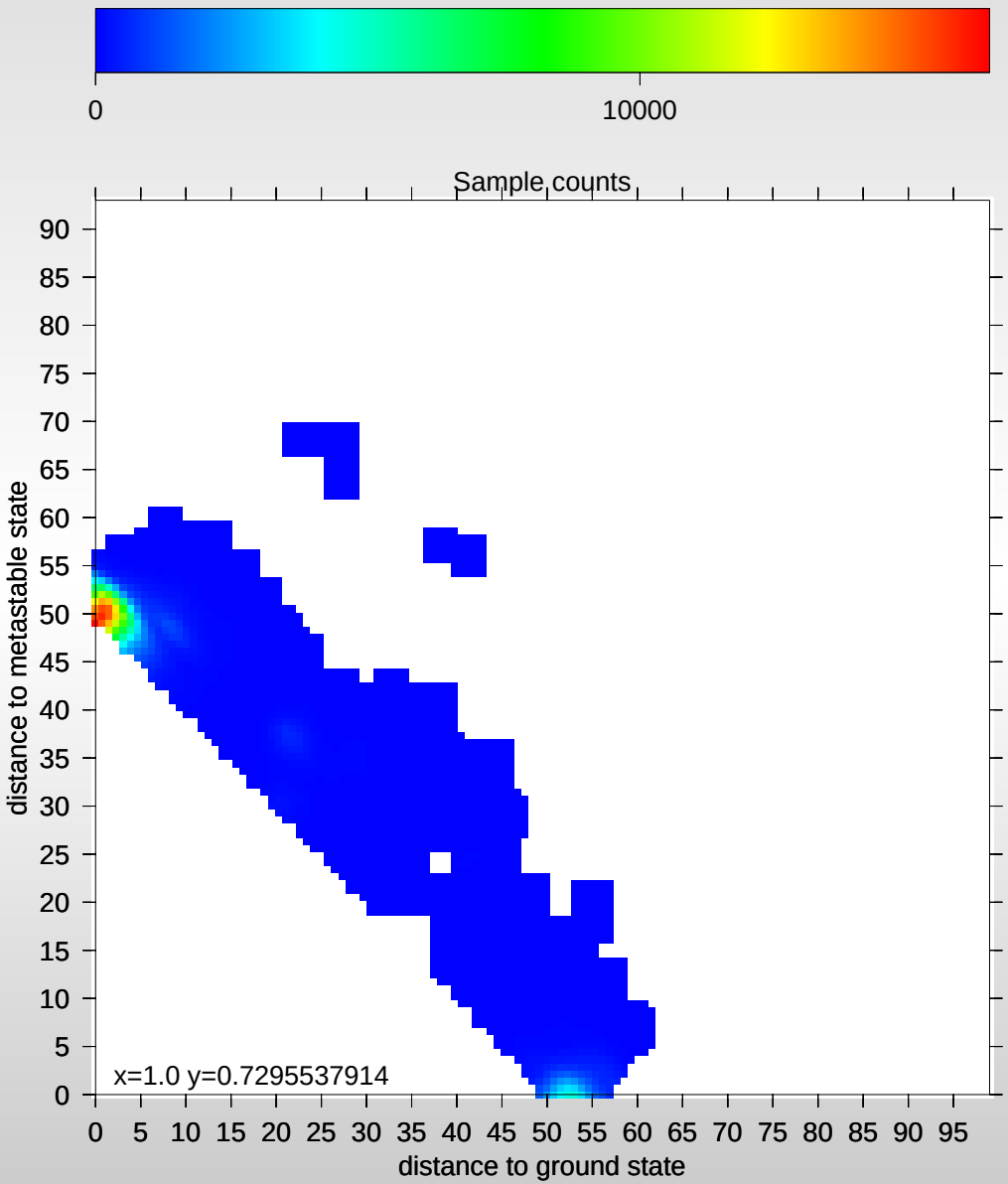
Example 3: 5'-UTR in MS2 - distortion 10^3



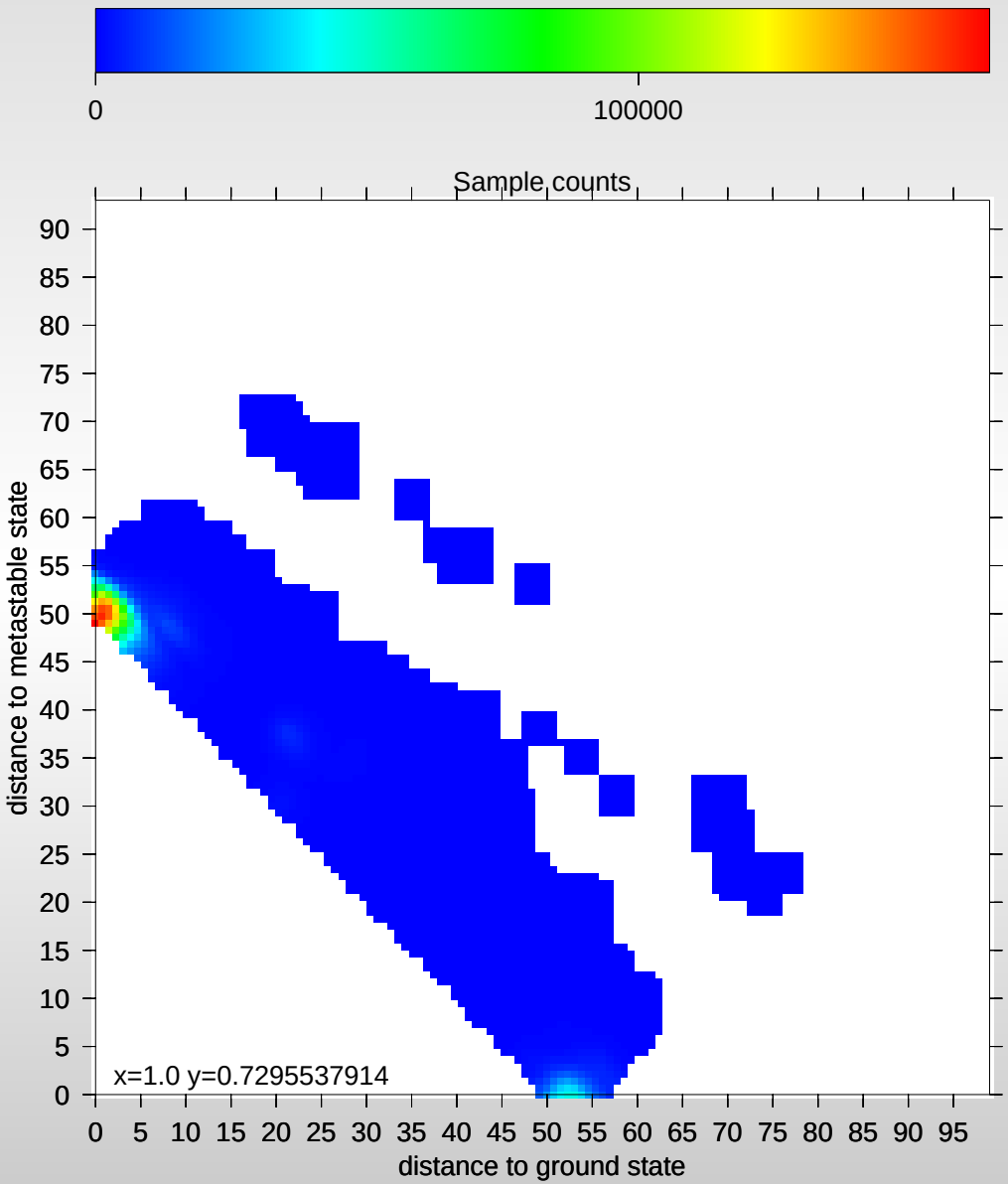
Example 3: 5'-UTR in MS2 - distortion 10^4



Example 3: 5'-UTR in MS2 - distortion 10^5



Example 3: 5’-UTR in MS2 - distortion 10^6



Still work in progress

- Exploration of the parameter space for the RNA2Dfold approximating soft constraint
- Construct adaptive algorithm to fill up distance classes with representatives

- Provide simpler variants of hard/soft constraints for executable programs within ViennaRNA
- Provide generalized variants for usage with RNAlib
- Reimplementation of RNApbfold such that it exploits the new features in ViennaRNA

Thanks to

- Yann Ponty
- Peter F Stadler
- Ivo L Hofacker

Thank You for your attention!