Sampling strategies to approximate RNA folding kinetics

Gregor Entzian

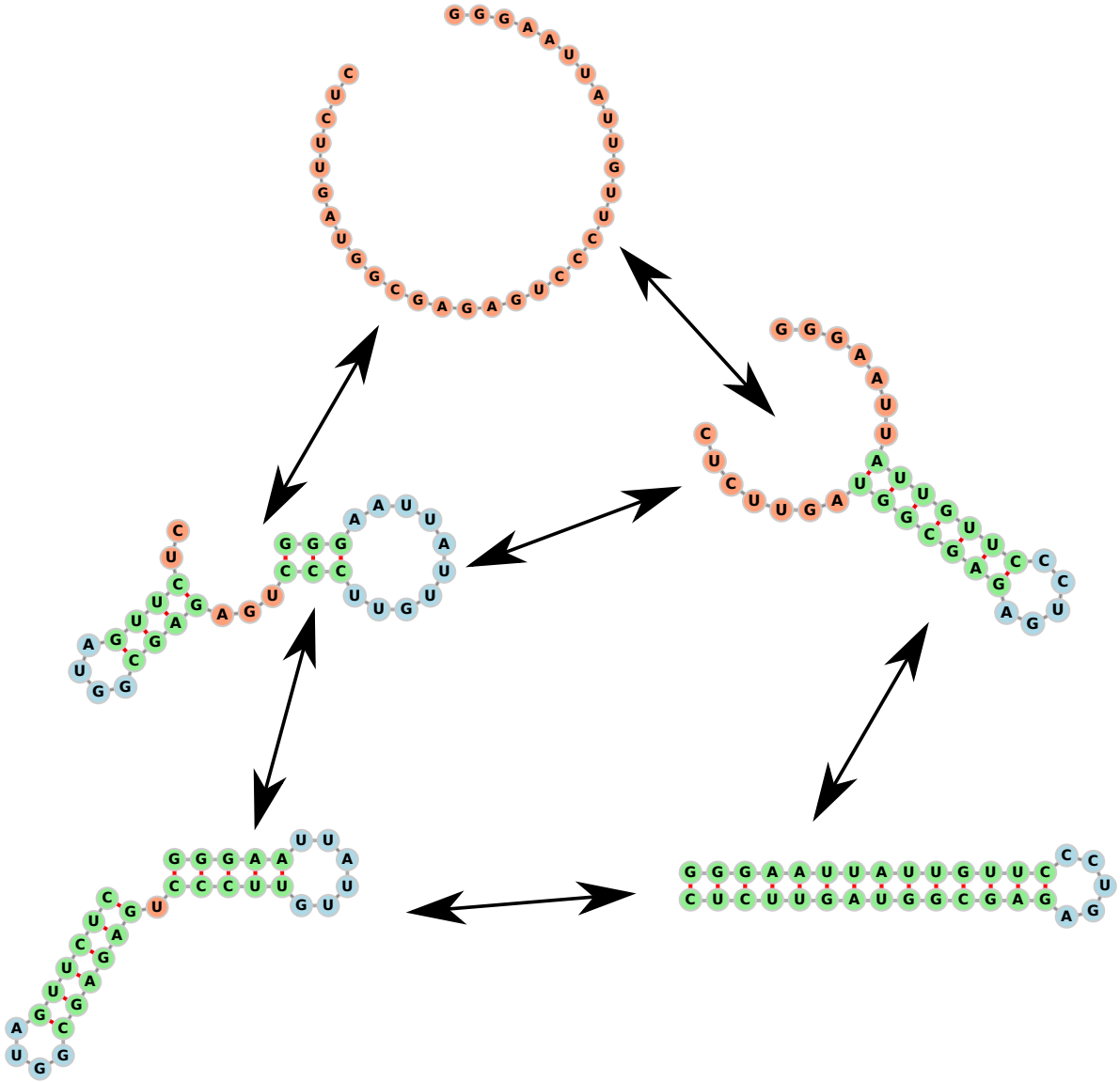
University of Vienna, Faculty of Chemistry, Department of Theoretical Chemistry

entzian@tbi.univie.ac.at

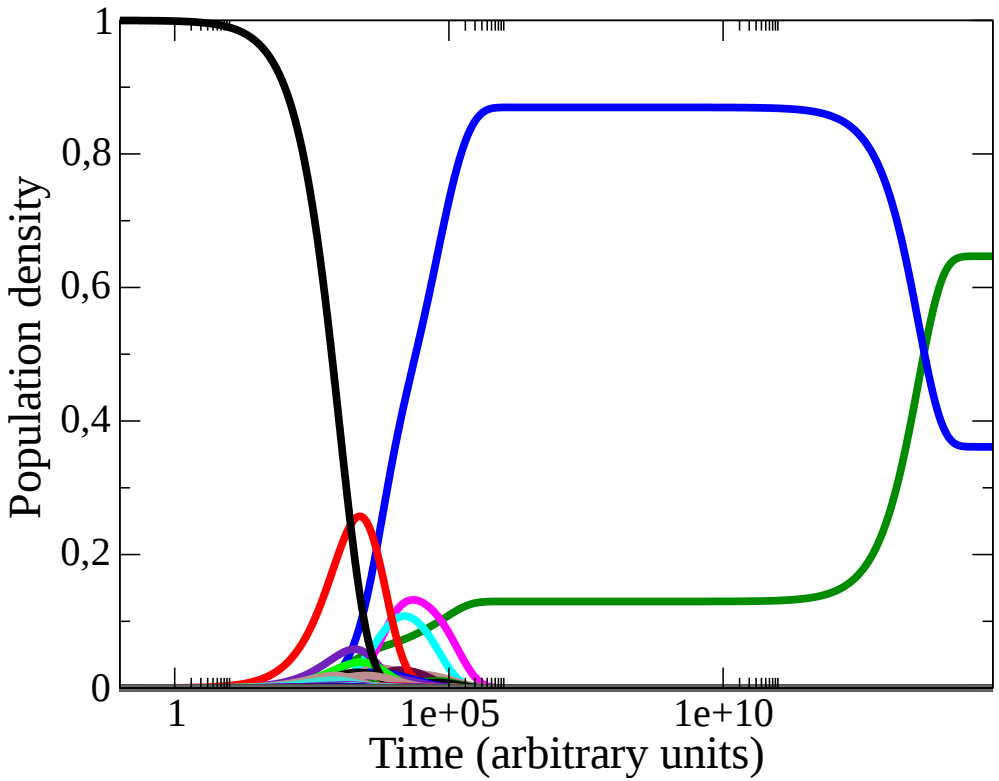
Supervisor: Univ.-Prof. Dipl.-Phys. Dr. Ivo L. Hofacker

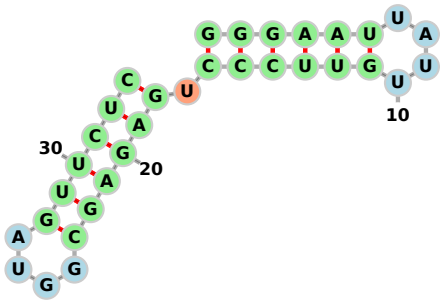
February 14, 2017

RNA Structure Dynamics

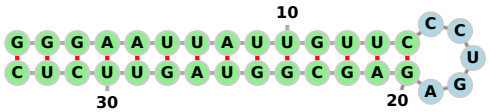


RNA Switch Folding Kinetics





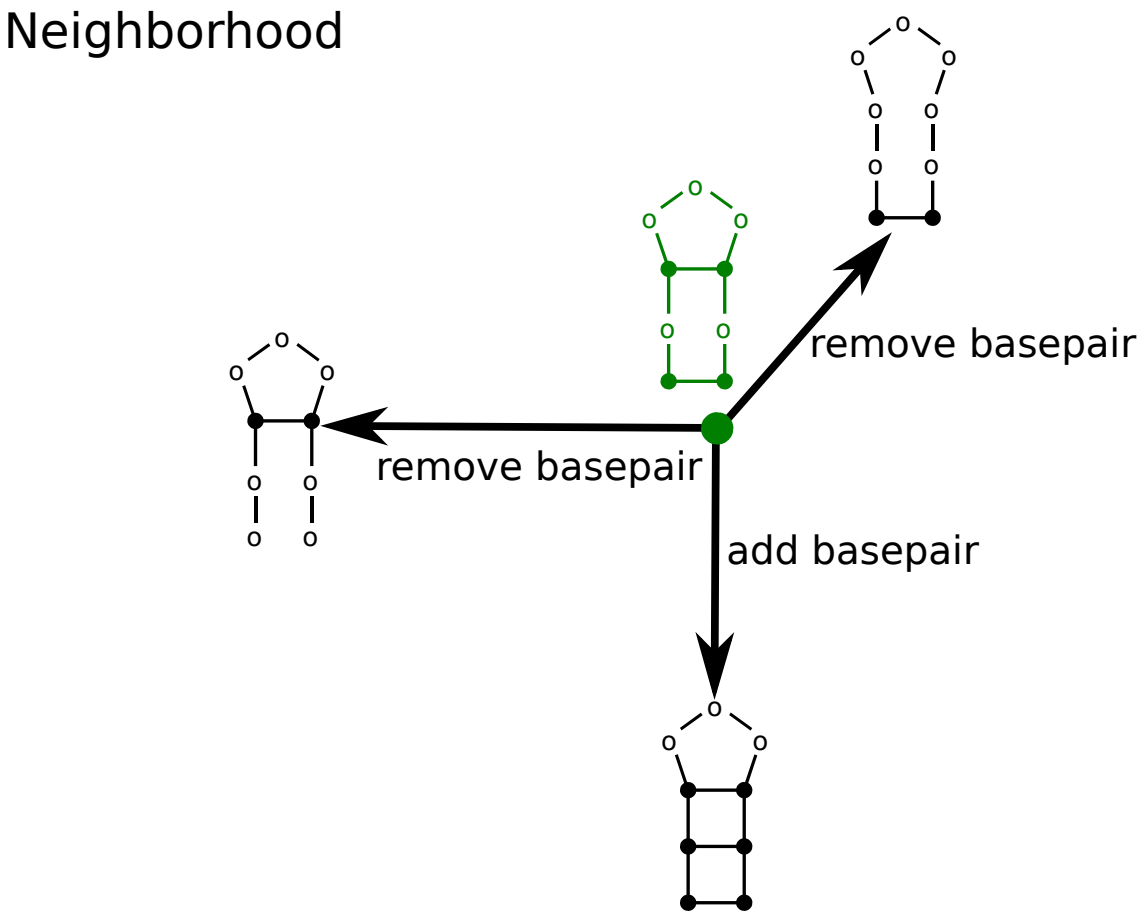
Ground state "On"



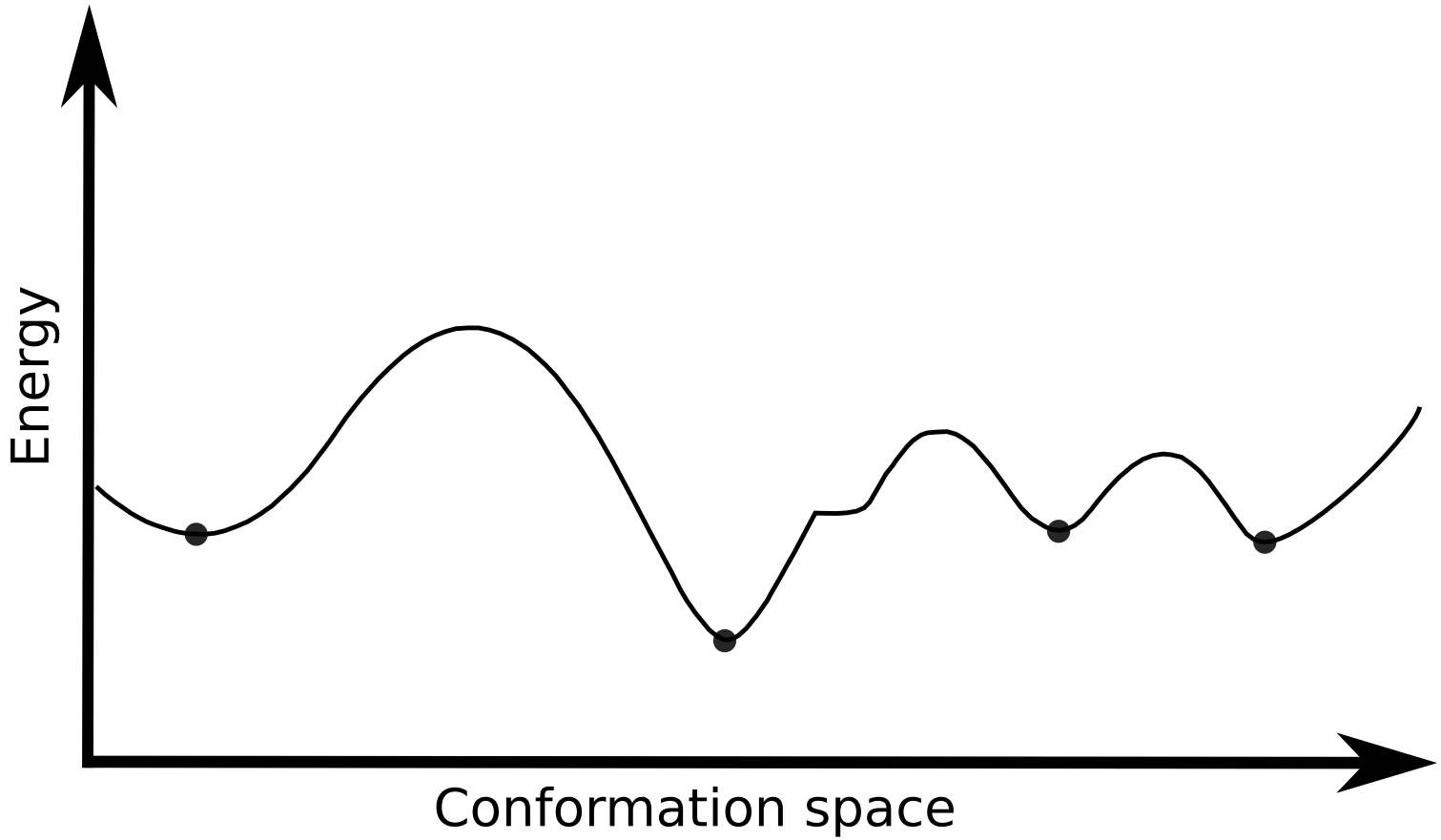
Alternative state "Off"

Energy Landscape

$$L = \{S, f, M\}; S = \{s_1, s_2, ...\}; f : s \to \mathbb{R}; \mathrm{M} = \text{" } + / - 1\mathrm{bp}\text{"}$$



Energy Landscape - Growth

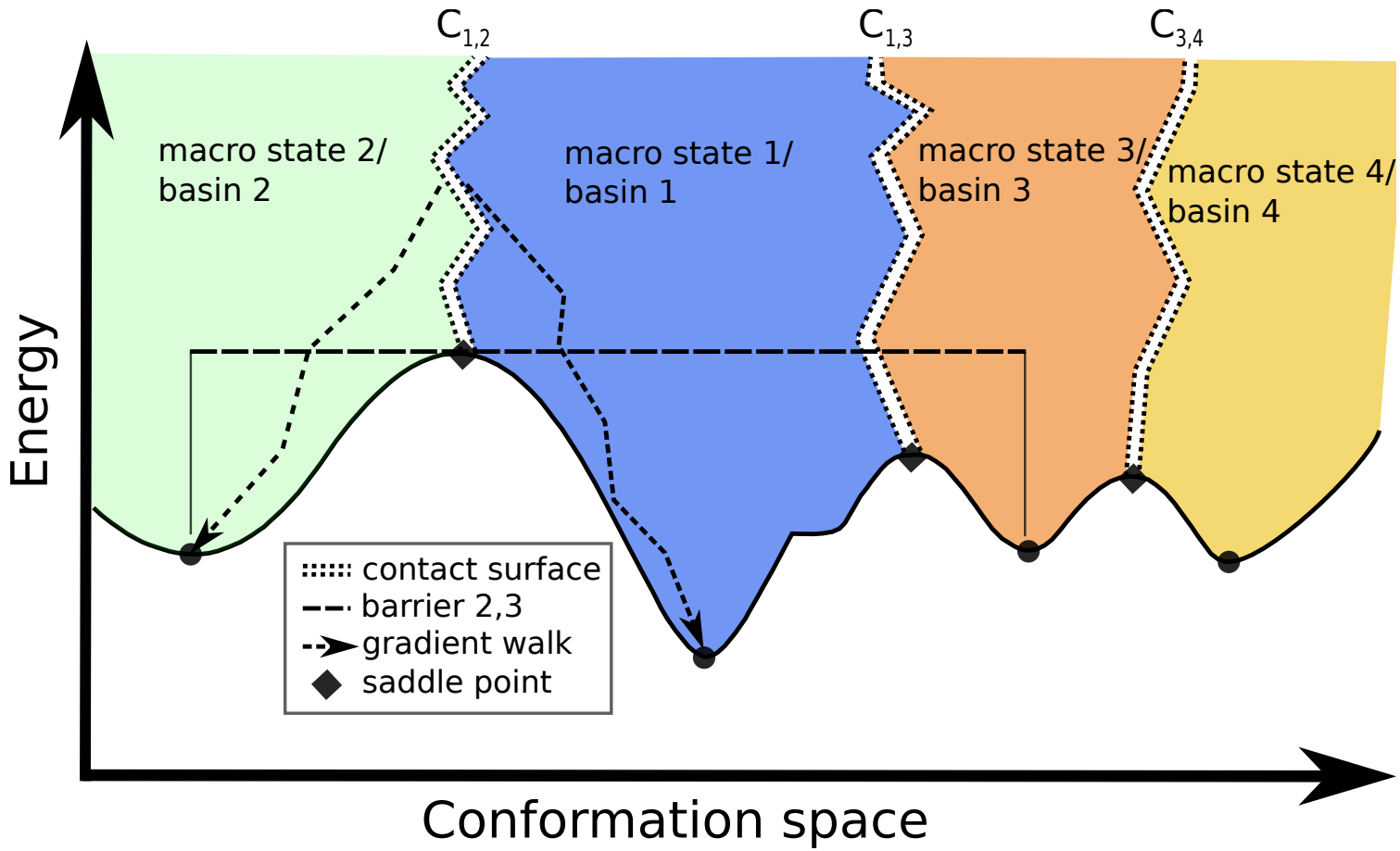


growth: 1.8^N

atoms in the universe: 10^{80}

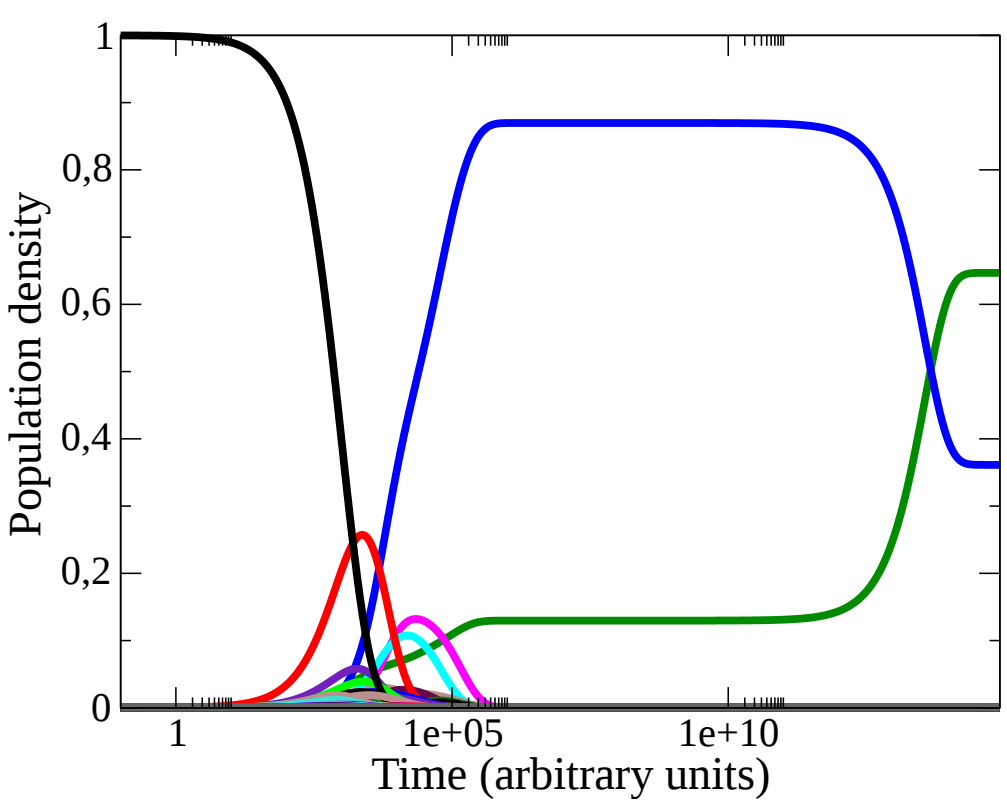
$1.8^N = 10^{80} \rightarrow N = 312 nucleotides$

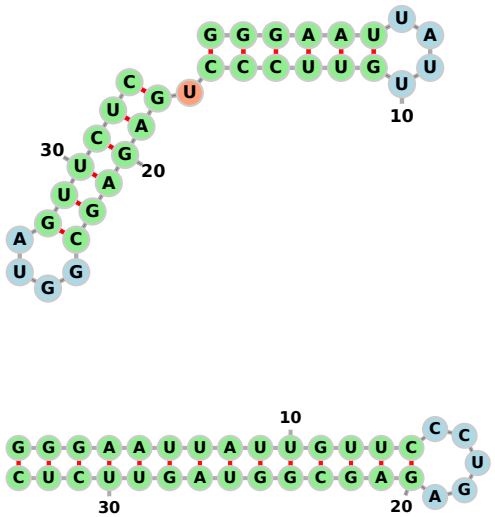
Coarse Graining



RNA Folding Kinetics

$$\frac{dp_i(t)}{dt} = \sum_{j \neq i} [p_j(t) r_{ji} - p_i(t) r_{ij}]$$



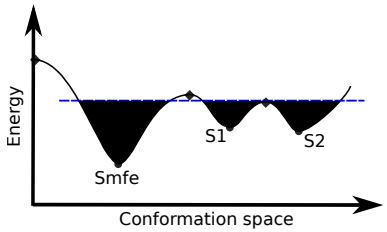


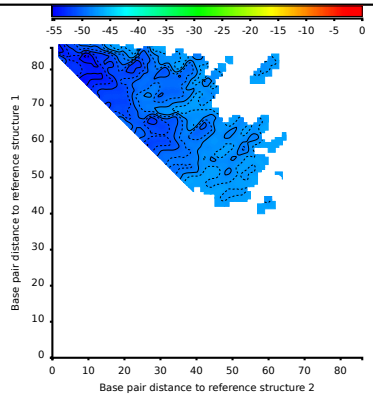
Goals

- Make RNA folding kinetics prediction applicable to biologically relevant sequence lengths ($> 100nt$)
- How to do that?
  1. Sample the energy landscape to cover most important states
  2. Determine physically meaningful partitioning of partial landscape into macro states
  3. Derive good transition rate approximations for resulting macro states
  4. Implement the above into a pipeline

State Generation

Exhaustive enumeration
- up to a certain threshold
- exponential number of structures
$\rightarrow$ only for short sequences

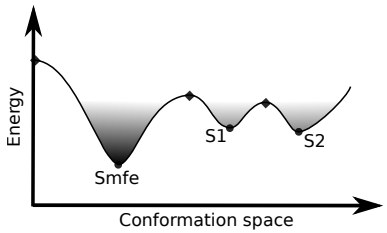


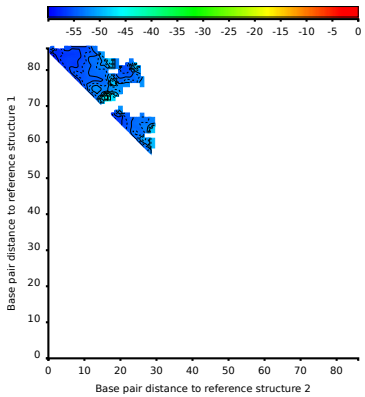


Boltzmann sampling
(importance sampling)

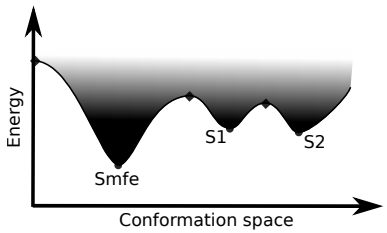
$P(s) = \frac{e^{-\frac{E(s)}{RT}}}{Q}$ with $Q = \sum_S e^{-\frac{E(s)}{RT}}$

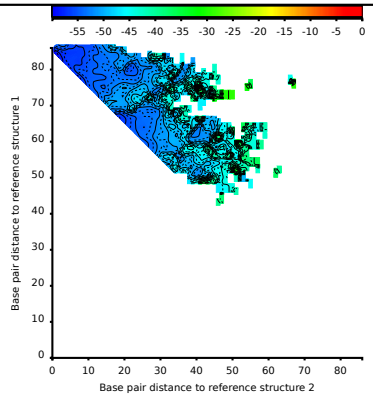
- highly redundant
- only structures with small energy deviations from the ground state





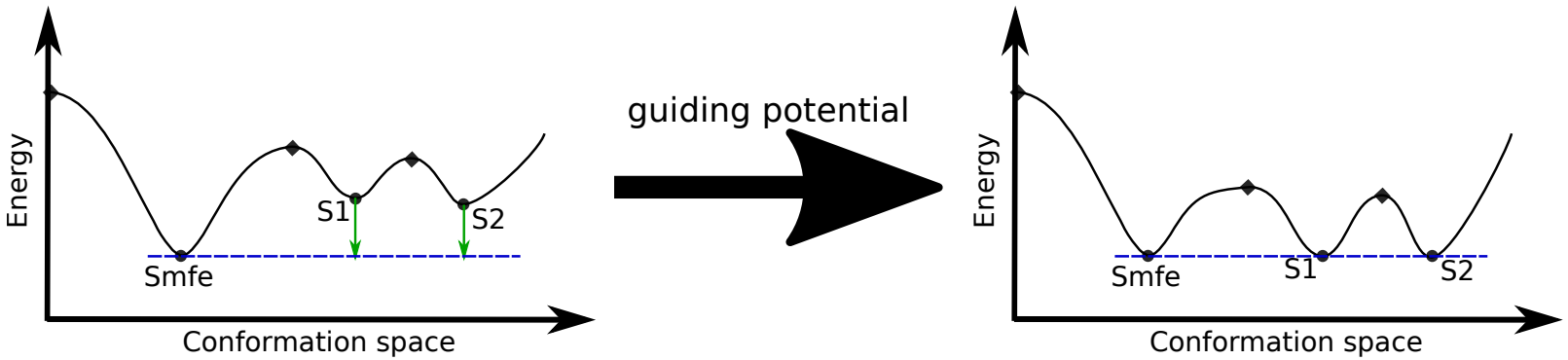
Variable temperature sampling
$T = \xi \cdot T_0, \xi > 1$
- undirected
- large T = sampling from uniform distribution





New Approach: Guiding Potentials

- sampling with focus on important reference structures



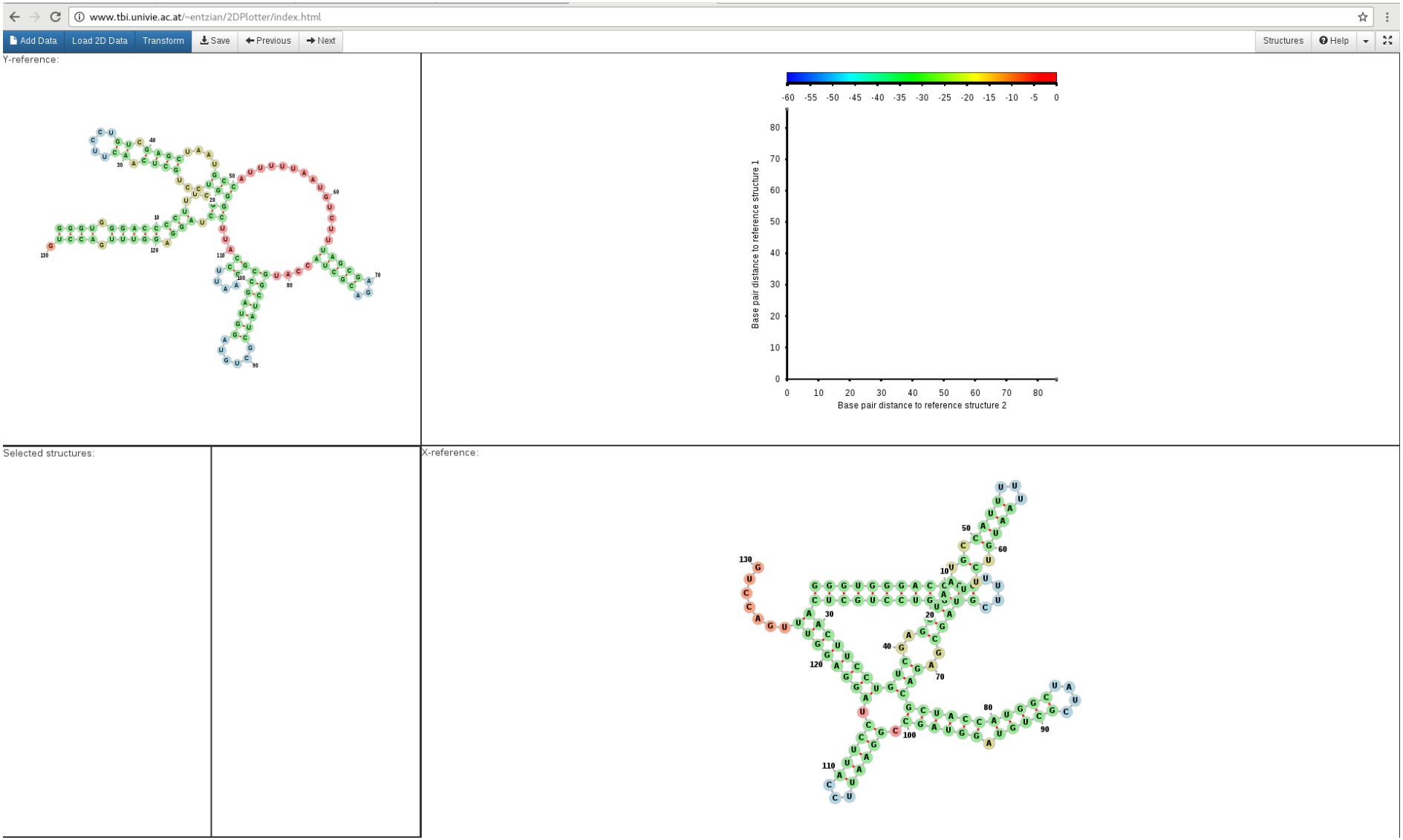
$$p(s_1) = p(s_2) = p(s_{mfe})$$

$$E'(s) = E(s) + \hat{E}(s)$$

$$\hat{E}(s) = d(s, s_1) \cdot w_1 + d(s, s_2) \cdot w_2$$

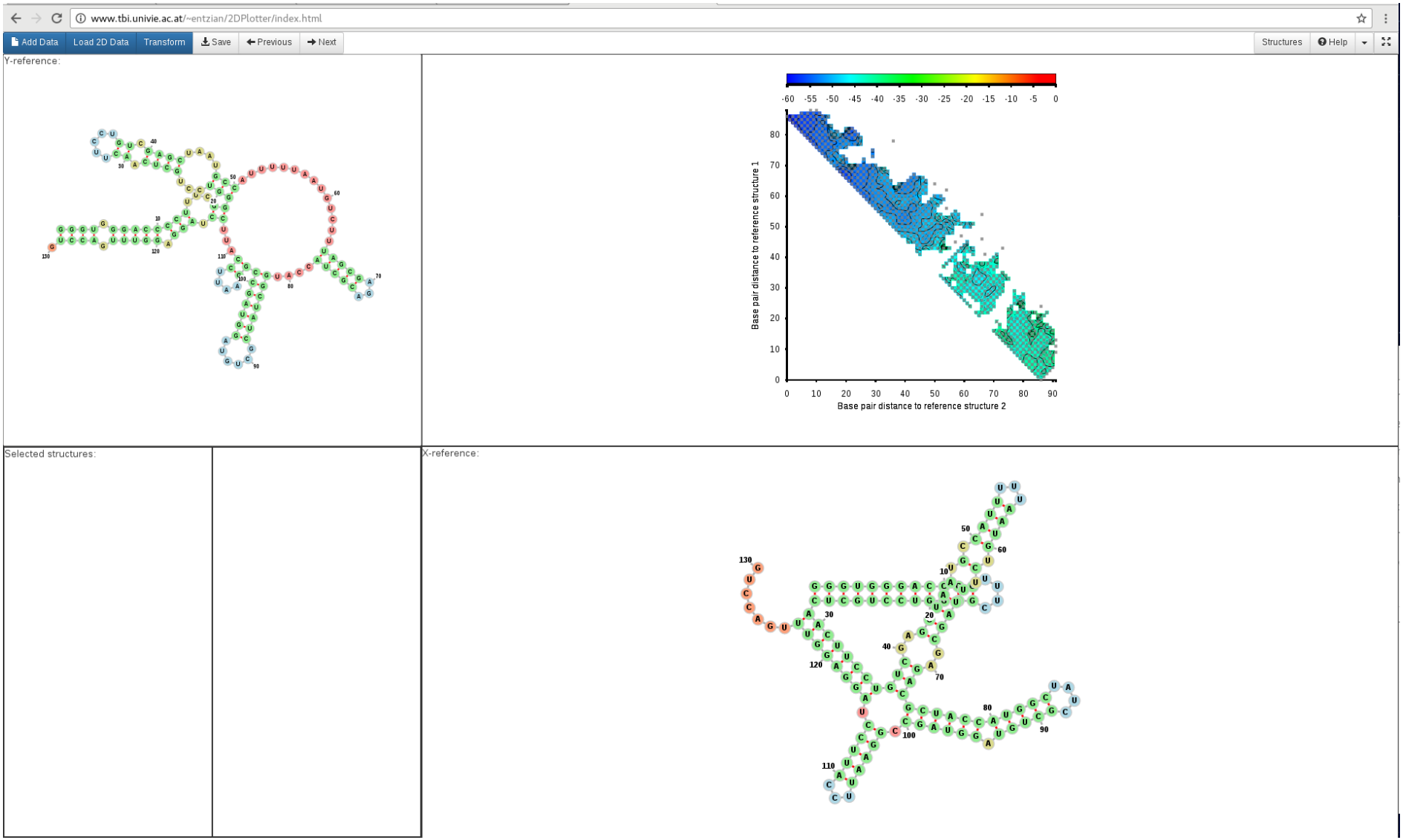
Interactive Guided Sampling - Initial References

`http://www.tbi.univie.ac.at/~entzian/2DPlotter`



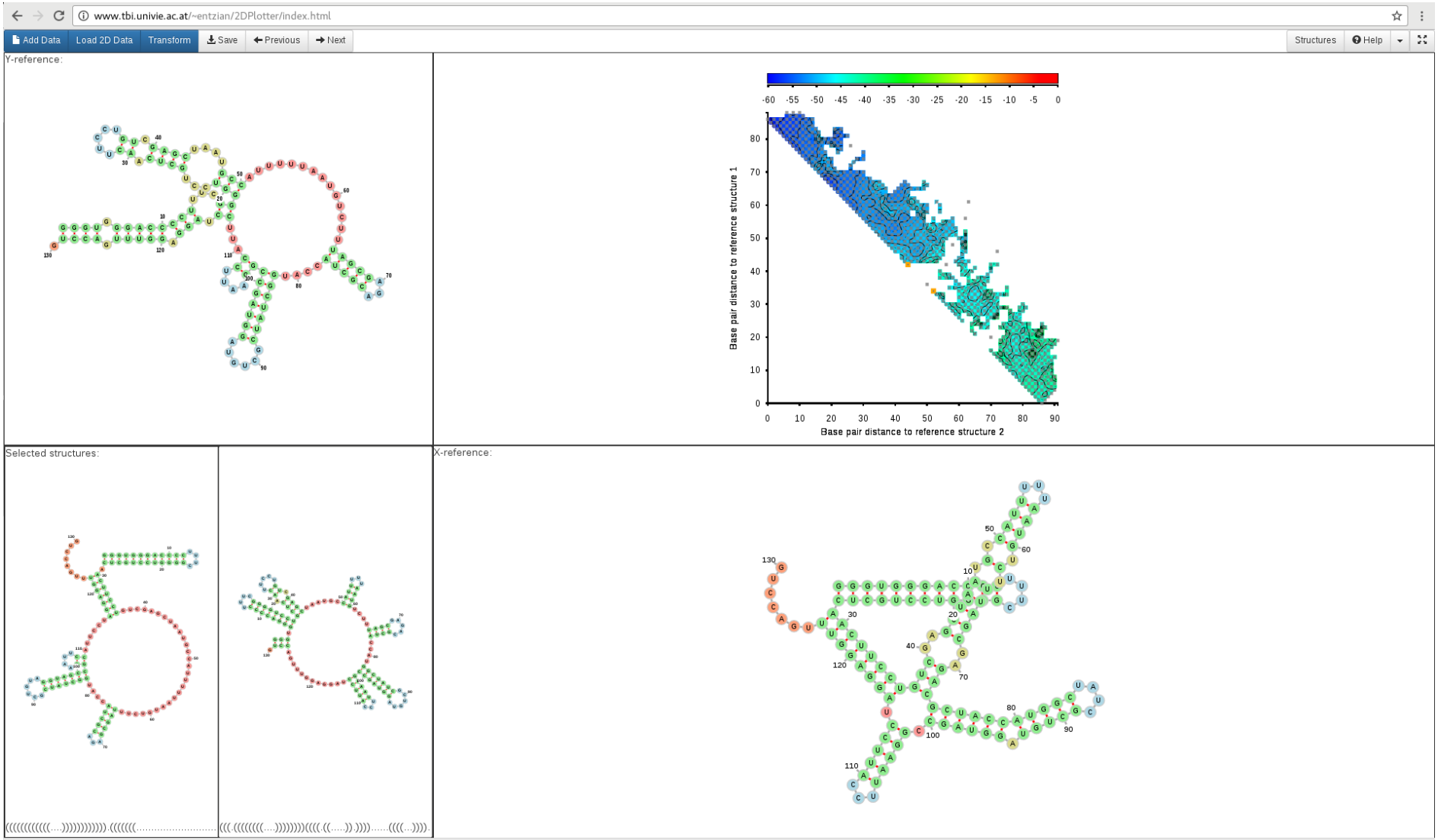
Interactive Guided Sampling - Iteration 1

`http://www.tbi.univie.ac.at/~entzian/2DPlotter`



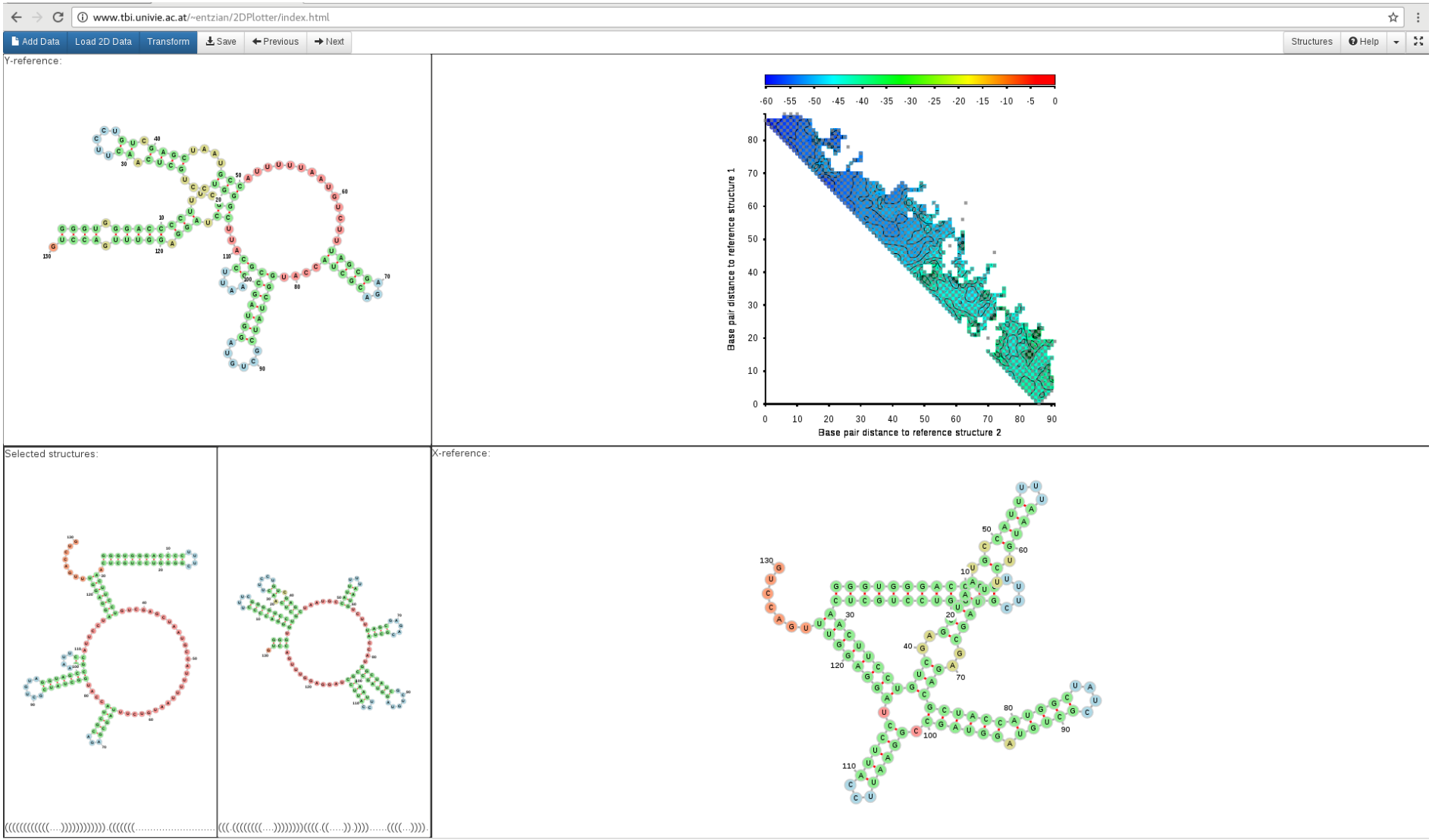
Interactive Guided Sampling - New References

`http://www.tbi.univie.ac.at/~entzian/2DPlotter`



Interactive Guided Sampling - Iteration 2

`http://www.tbi.univie.ac.at/~entzian/2DPlotter`



Tasks

1. Develop an automated iterative **sampling strategy** which is fast and produces both, diverse and most important structures of the energy landscape.
2. Develop **cluster-strategies** for RNA structures, to identify important structures and to generate macrostates.
3. Develop methods for computing the **transition rates** for incomplete landscapes.
4. Construct a **pipeline and programs** to compute RNA folding kinetics for long sequences and explore the underlying energy landscape. Implement a **web server**, which provides a comfortable graphical user interface.

Thank you!

- Ivo Hofacker
- RNALands project team:
  - Andrea Tanzer, Ronny Lorenz, Maria Waldl, Yann Ponty, Mireille Regnier, Hélène Touzet, Loic Paulevé, Alain Denise, Juraj Michalik