Conservation as Soft Constraint

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,
University of Leipzig
Max Planck Institute for Mathematics in the Sciences
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
Center for non-coding RNA in Technology and Health, U. Copenhagen
The Santa Fe Institute (external faculty)

Universidad Nacional de Colombia (prof. hon.)

joint work with the Hofacker Lab in Vienna

Benasque, Jul 2024

Features and Soft Constraints

- A feature μ is simply a subset of secondary structures that have something in common.
- $\mathbb{P}(\mu) = \sum_{z \in \mu} exp(-E(z)/RT)/Z$
- External evidence for a feature μ, quantified as a probability $p[\mu]$
- Usual version of a bonus energy

$$\Gamma_\mu = -RT \ln \frac{p[\mu]}{p[\neg\mu]}$$

- Dominating features: $p[\mu] > 1/2$. Example: centroid base pairs.
- Bonus energy $\Gamma_\mu < 0$ if and only if μ is dominating

Consensus folding using `RNAalifold`

- `RNAalifold` uses the same algorithms and energy parameters as `RNAfold`
- Energy contributions of the single sequences are averaged
- Covariance information (e.g. compensatory mutations) is incorporated in the energy model.
- It calculates a consensus MFE consisting of an energy term and a covariance term:

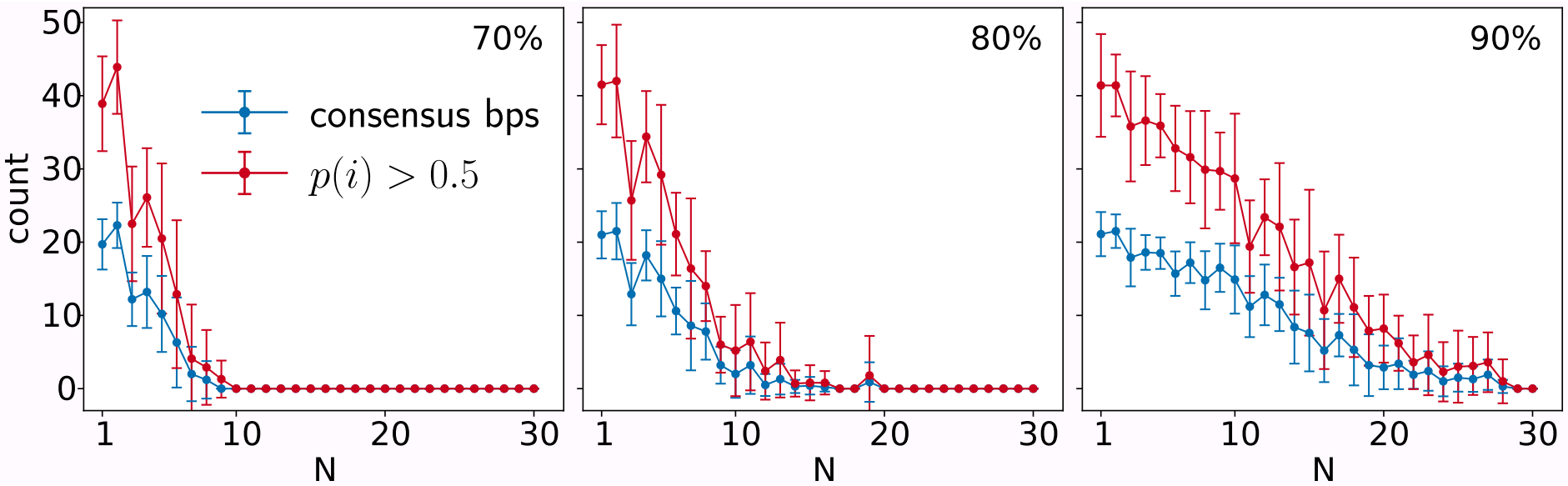
```
((((((..(((.........))).(((((.......)))))).....(((((.......)))))))))).
GTTTCCGTAGTGTAGCGGTTATCACATTCGCCTCACACGCGAAAGGTCCCCGGTTCGATCCCGGGCGGAAACA
GTTTCCGTAGTGTAGTGGTTATCACGTTCGCCTAACACGCGAAAGGTCCCCGGTTCGAAACCGGGCGGAAACA
GTTTTCGTAGTGTAGTGGTTATCACGTGTGCTTCACACGCACAAGGTCCCCGGTTCGAACCCGGGCGAAAACA
**** ********** ********* *  ** * ******  ****************  ******* *****
(-24.76 = -23.43 +  -1.33)
```

A partition function version also computes the base pairing probabilities and marginal probabilities $p(i)$ that position i is paired.

J.Mol.Biol. **319**:1059-1066 (2002)

Consensus Base Pairs

Alignments of *N* sequences with *x*% pairwise sequence identity (randomly placed mutations) have very few base pairs in common:



Absense of evidence (for conserved base pairs) $\neq$ evidence for unpaired bases!

ONLY use dominating consensus base pairs

= centroid base pairs of `RNAalifold`

RNAsoftcons

- Prepare an *sequence alignment* $\mathbb{A}$ that contains the focal sequence x.
- Compute the partition function of the alignment with `RNAalifold` and extract:
  - **a** base pairs with $p(i,j) > 1/2 \rightarrow \Gamma_{i,j} = RT \log \frac{p(i,j)}{1-p(i,j)}$
  - **b** significantly paired positions $p(i) > 1/2 \rightarrow \Gamma_{i,j} = RT \log \frac{p(i)}{1-p(i)}$
- Project these "centroid consensus" bonus energies onto the focal sequence x
- Fold x with `RNAfold` with the bonus energies as "soft constraint"

Performance

Data set: curated Rfam alignments project to *E.coli*.

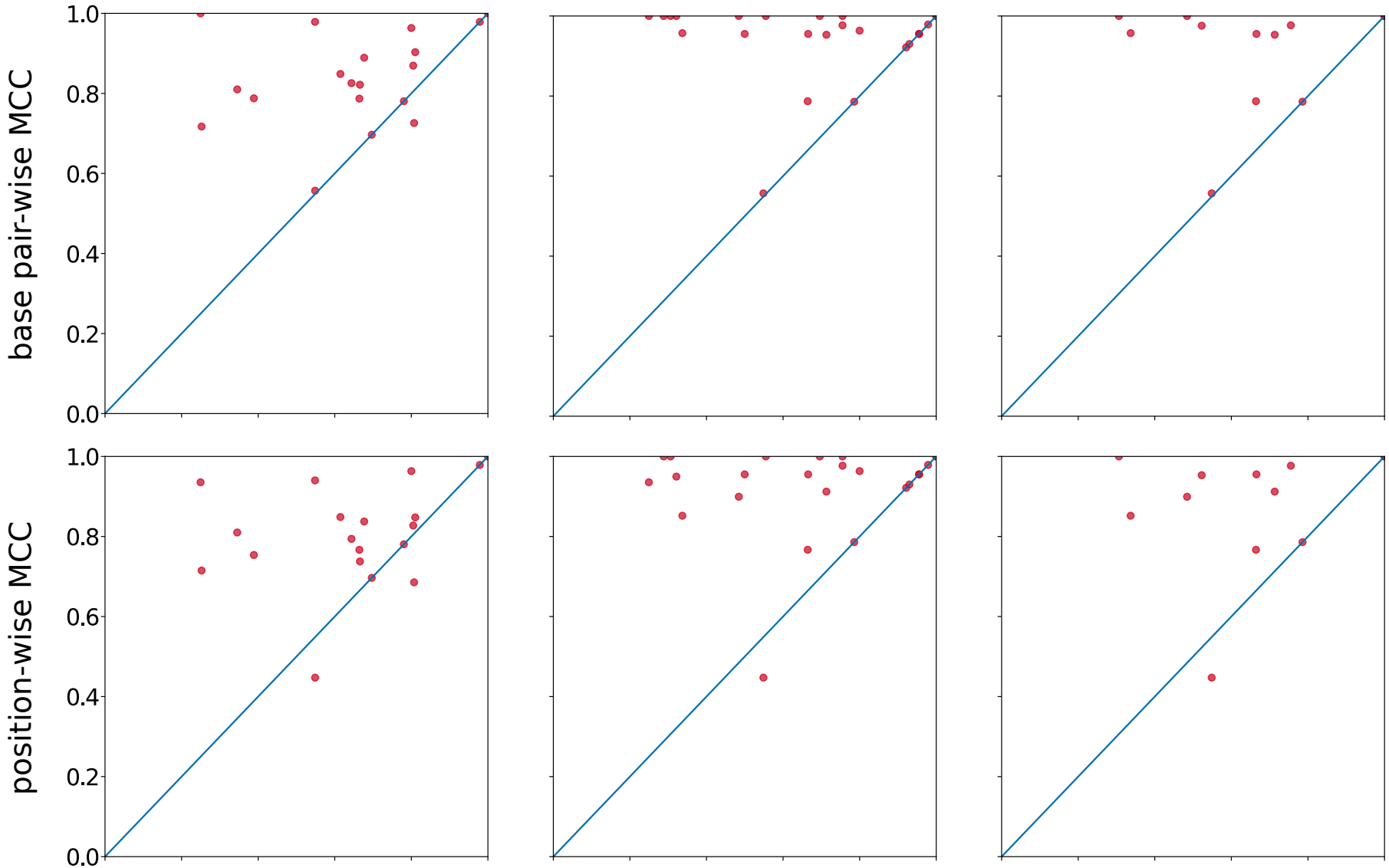
Method	MCC	F-val	PPV	Sensitivity
`RNAfold`	0.662	0.666	0.629	0.703
`refold.pl`	0.936	0.937	0.919	0.955
`RNAsoftcons`	**0.948**	**0.949**	0.922	**0.977**
`TurboFold`	0.931	0.932	**0.923**	0.940
`PETfold`	0.936	0.937	0.921	0.954

Base-pairs versus pairedness

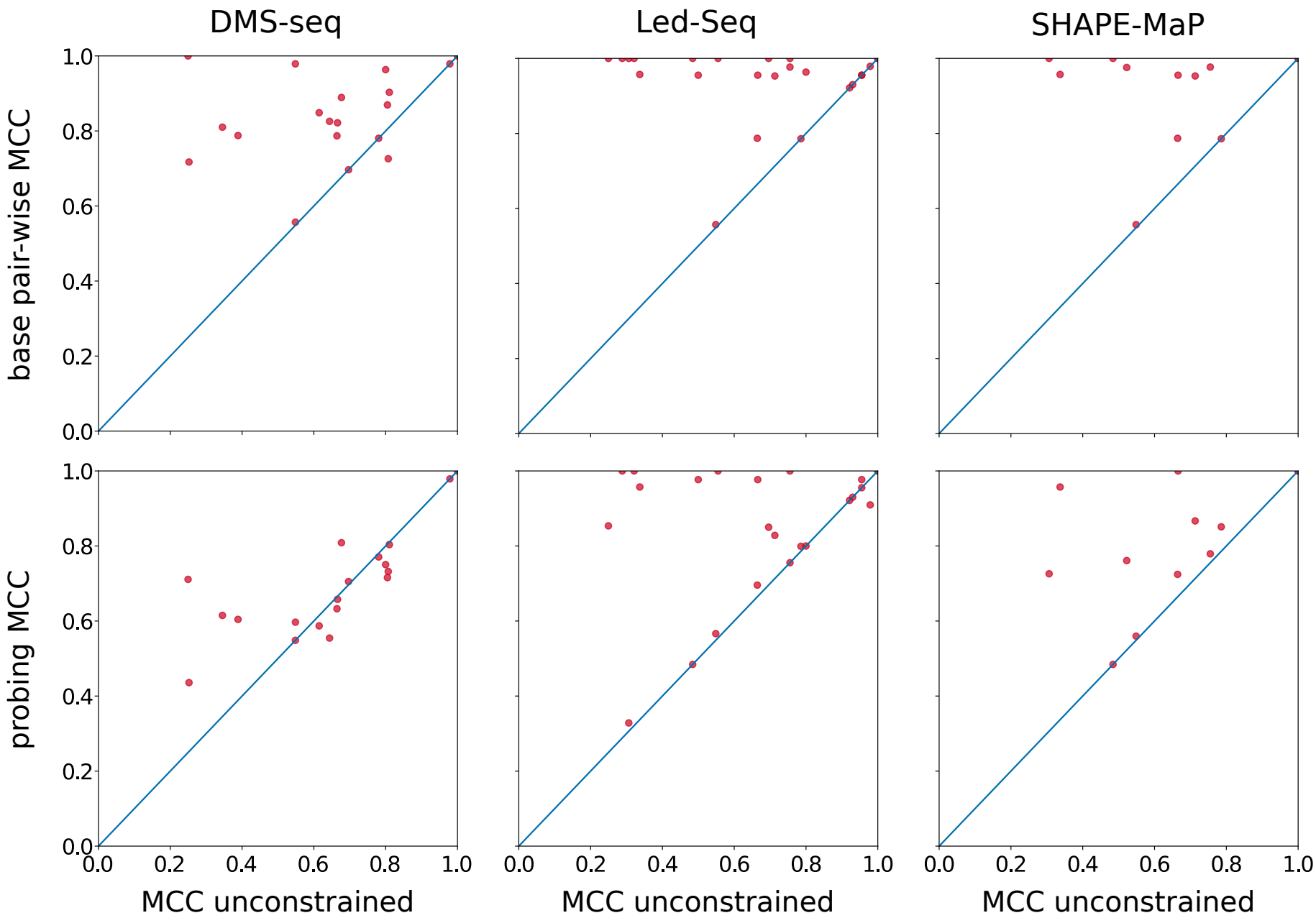
```
GCGCGATTAACGCGCTATGCGGGAAACCCGCGATTACGCGC
((((.....))))...((((...))))((....))))))     -1
((((.....((((...))((...))))))....))))))     -8
```

Same pairedness but very different base pairs!

Comparison in accuracy

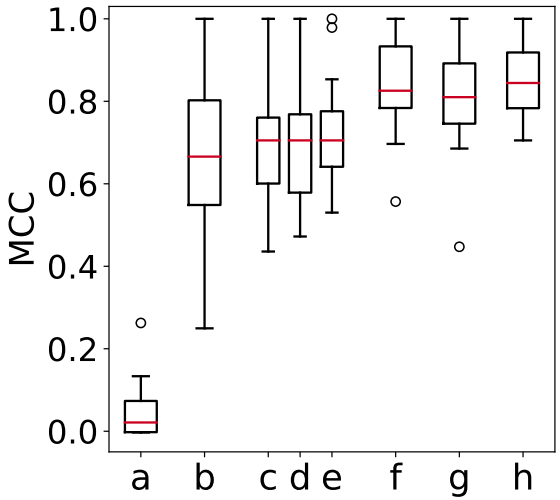


Comparison: Base pairs versus probing

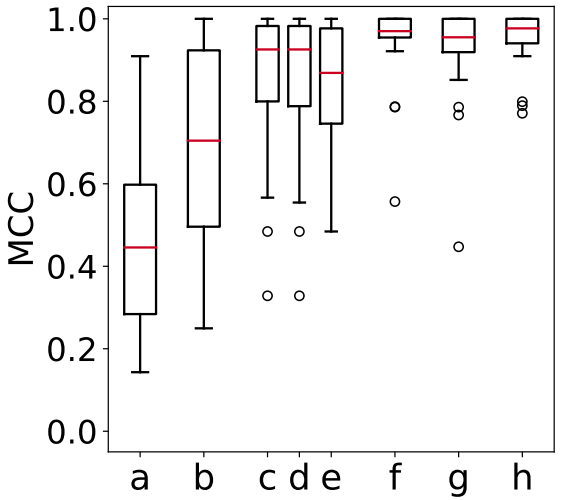


Probing versus Conservation

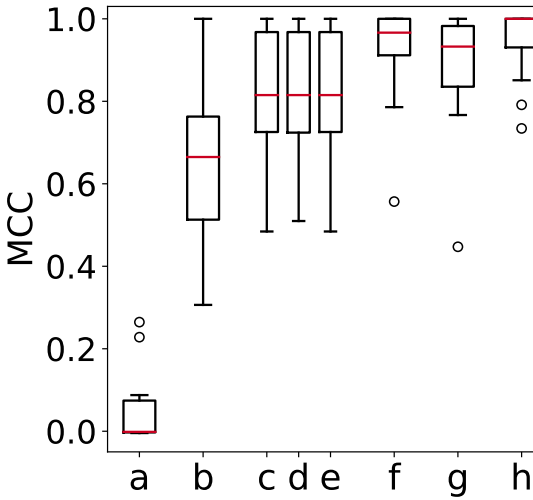
DMS-seq



Led-Seq



SHAPE-Map



a: chemical probing data alone
b: thermodynamic folding with Turner energy model (`RNAfold`)
c, **d**, **e**: Chemical probing data used as pseudo-energies in thermodynamic folding: (three different ways of converting signal to pseudo-energy)
f, **g**: Phylogenetic information used a pseudo-energies in thermodynamic folding: **f** base pair-wise information **g** position-wise information only
h Combination of chemical probing and phylogenetic information.

Conclusion

(1) Conservation data work well as “soft constraints” to obtain much improved secondary structure predictions
— if there is a structural consensus

(2) Probing data alone, i.e., without the Turner energy model (or SCFG equivalent) convey very little information

(3) In in doubt, conservation seems to give the better structure predictions, even when only pairedness is used!

Acknowledgements

Collaborators

- Sarah von Löhneysen (Leipzig)
- Ivo Hofacker's lab in Vienna:
  (T. Spicher, J. Varenyk, Hua-Ting Yao, Ronny Lorenz)

Funding

German Research Foundation (DFG), German Federal Ministery of Science (BMBF))