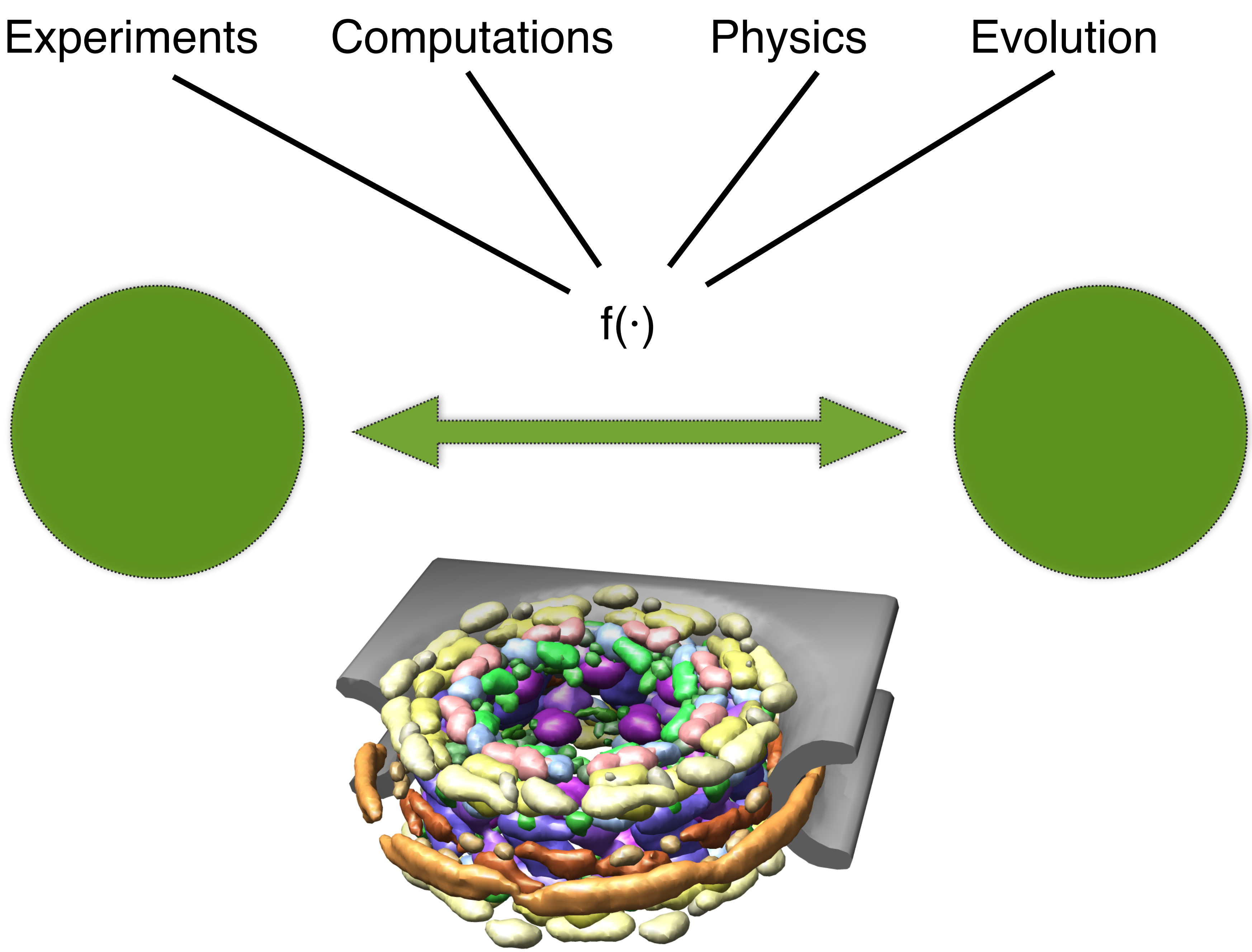
Integrative modeling applied to chromatin

François Serra, Marco Di Stefano & Marc A. Marti-Renom
Structural Genomics Group (CNAG-CRG)

The Integrative Modeling Platform

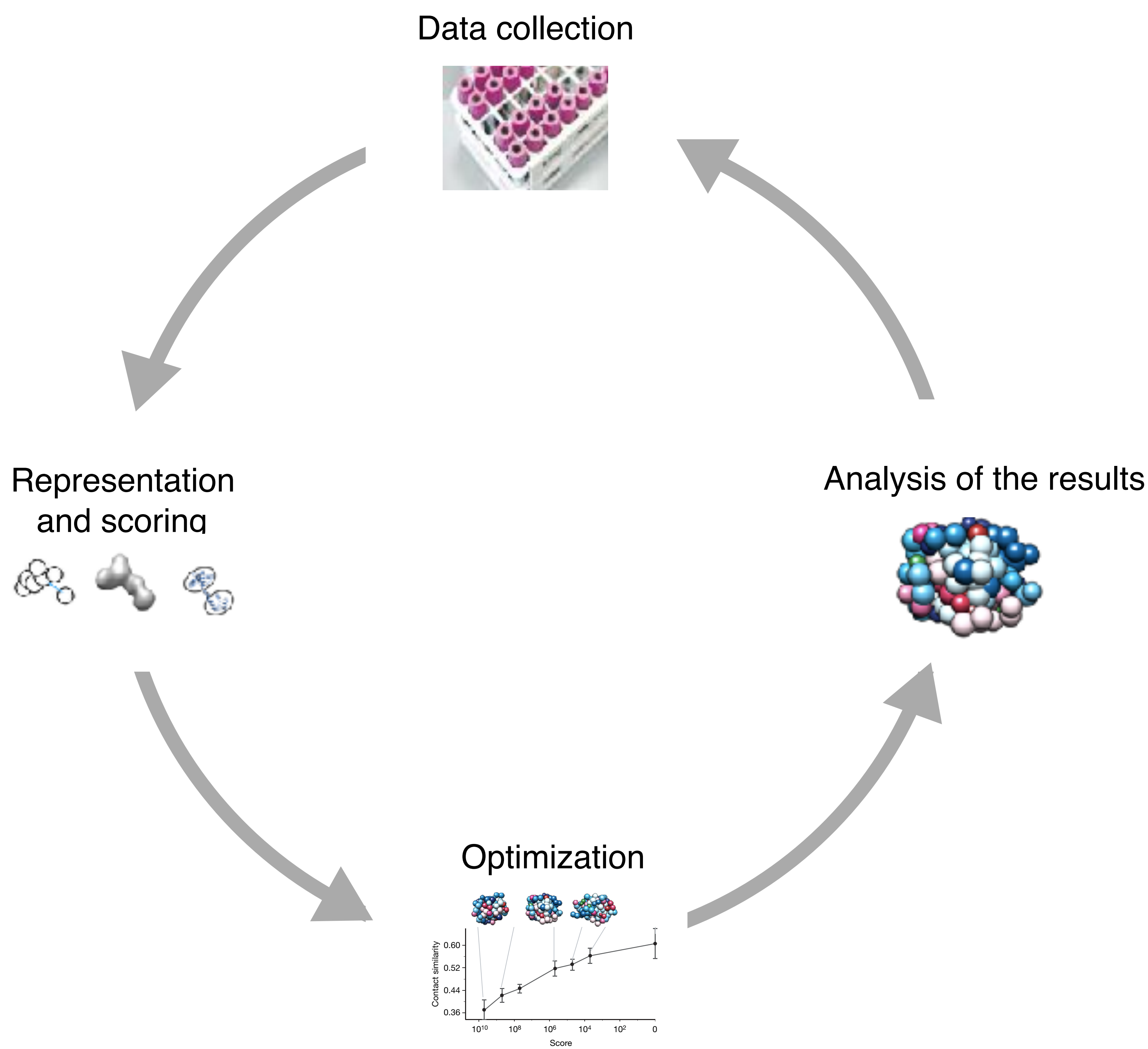
http://www.integrativemodeling.org

Russel, D. et al. PLOS Biology 10, e1001244 (2012).

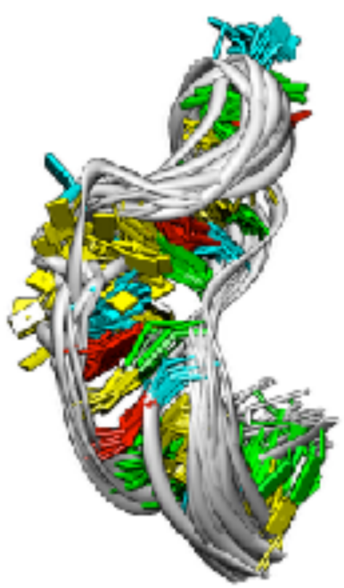


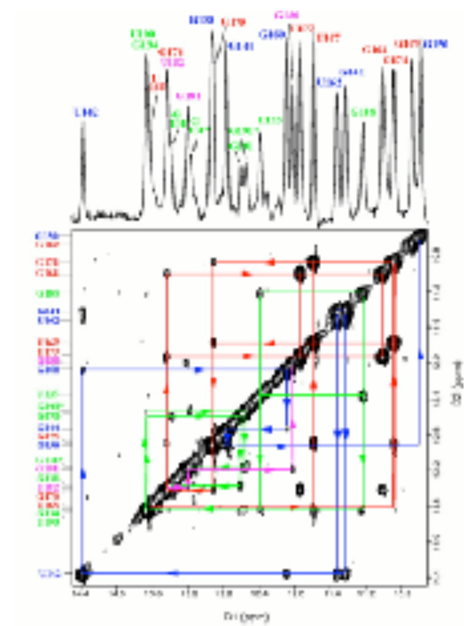
From Alber, F. et al. Nature 450, 695–701 (2007).

The four stages of integrative modeling

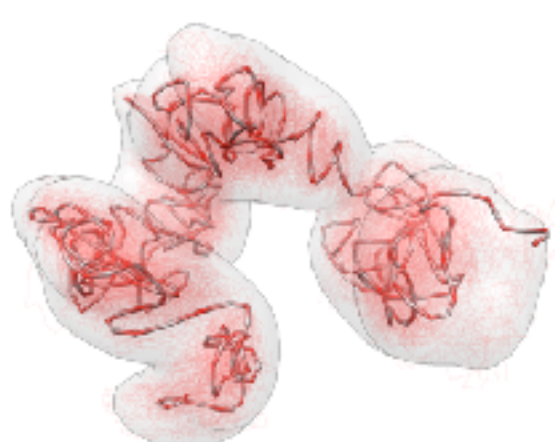


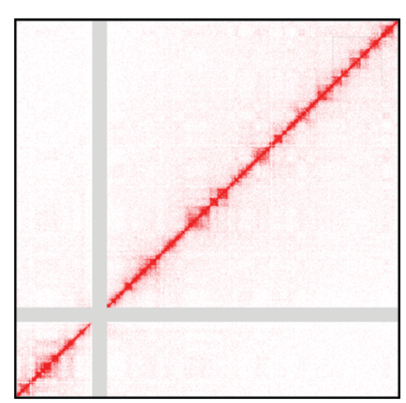
Structure determination using Hi-C data





Biomolecular structure determination
2D-NOESY data

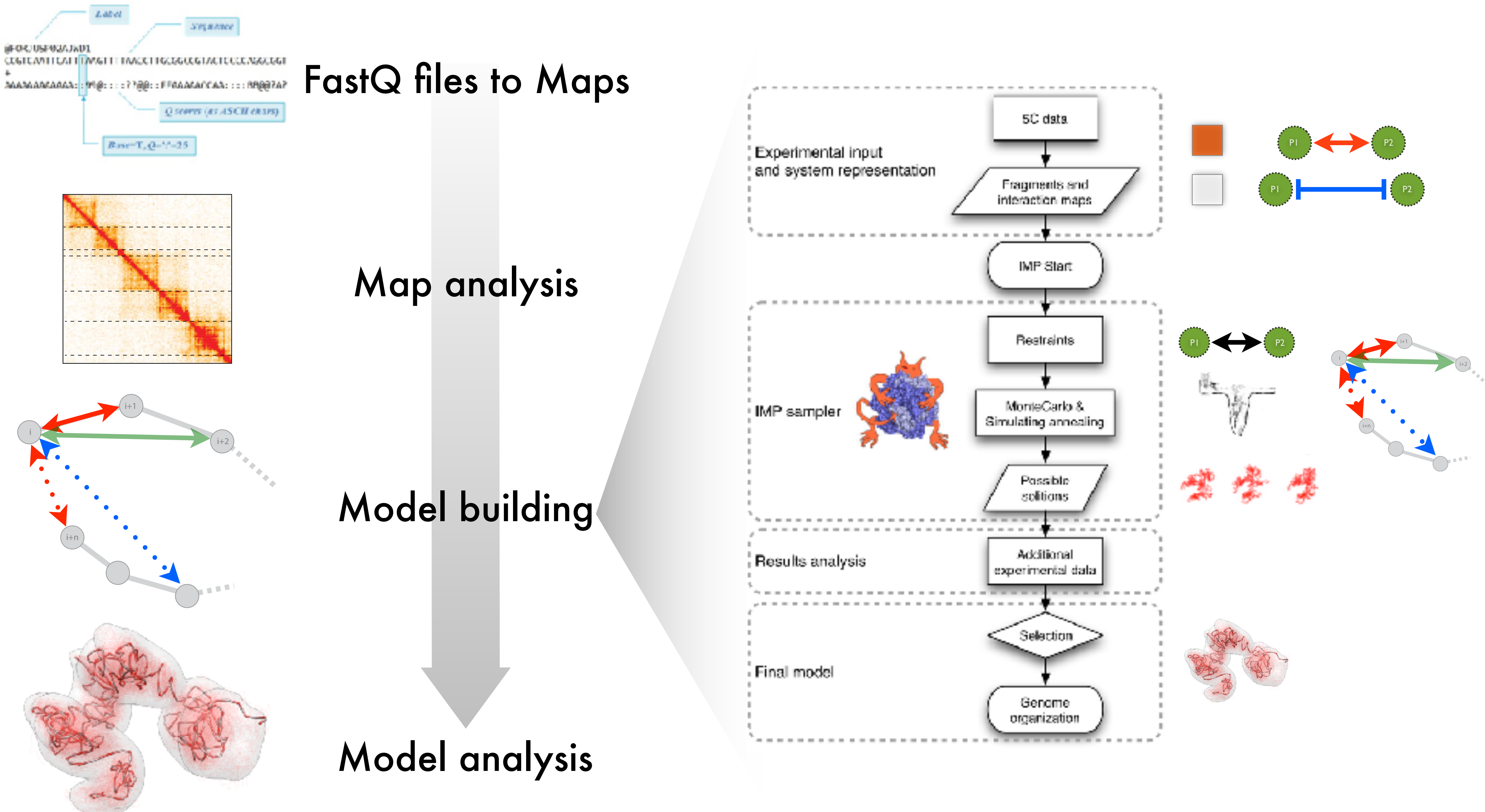




Chromosome structure determination
3C-based data

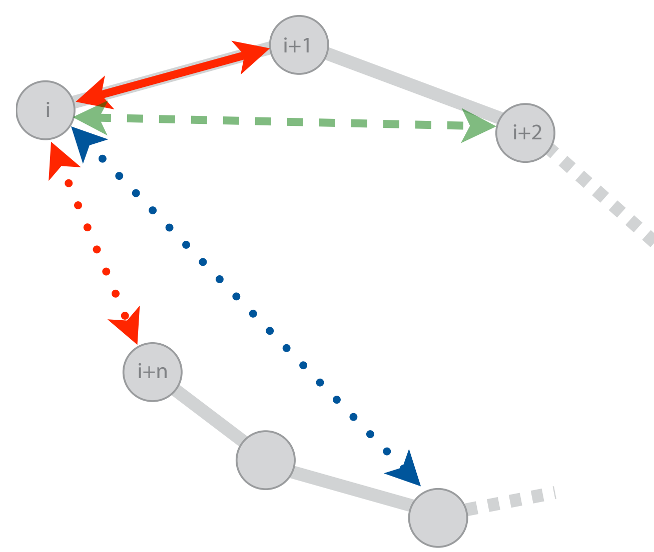
TADbit

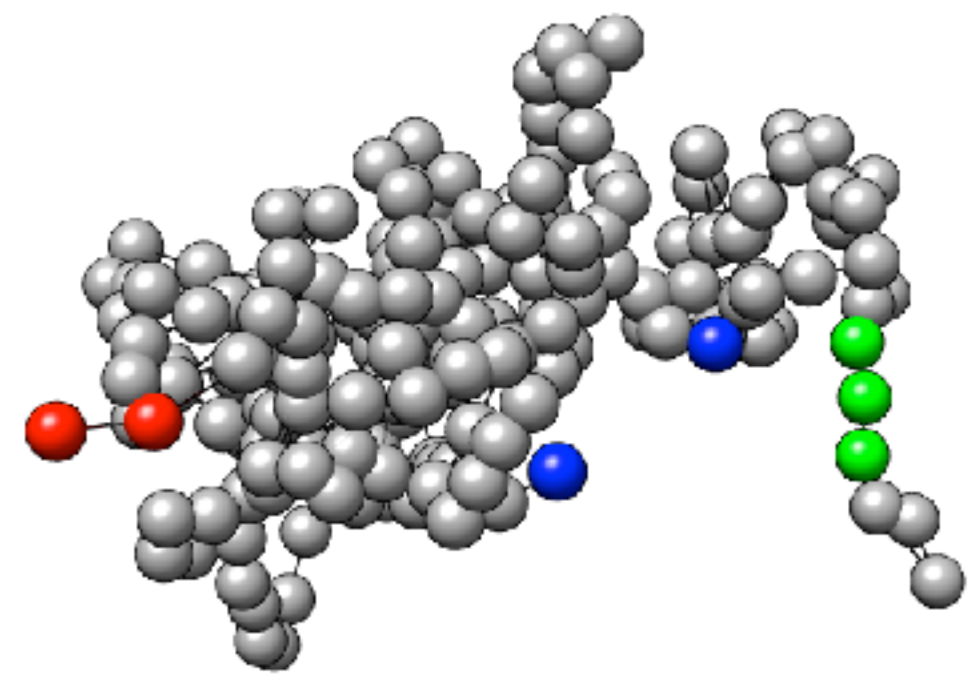
http://3DGenomes.org
http://www.integrativemodeling.org



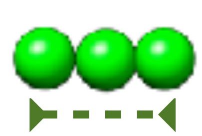
Model representation and scoring

Constituent parts of the molecule

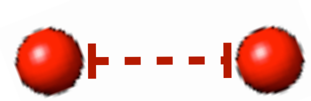




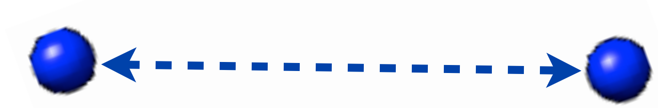
$d < d_0$



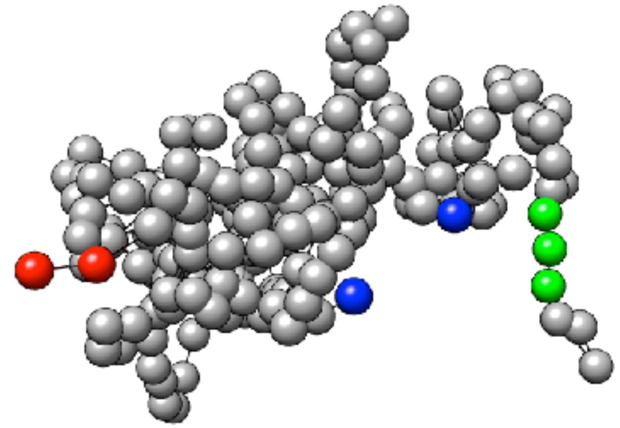
$d = d_0$



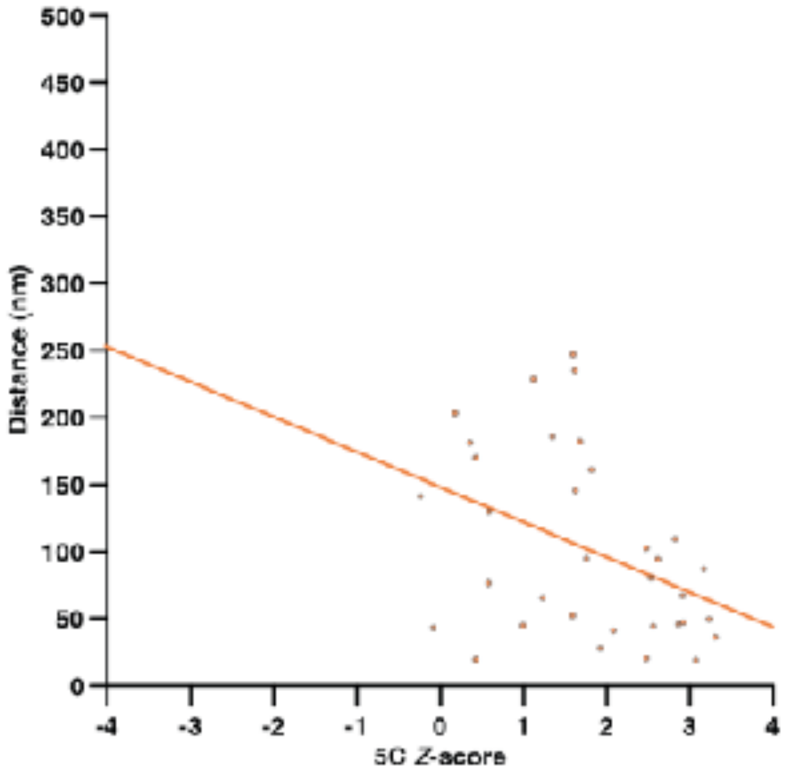
$d > d_0$



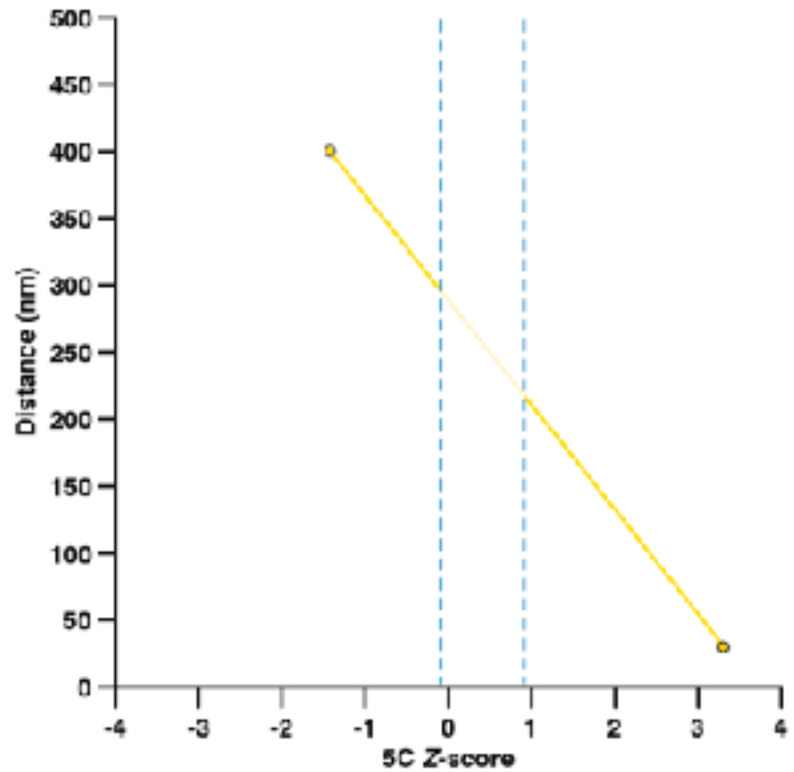
From 3C data to spatial distances



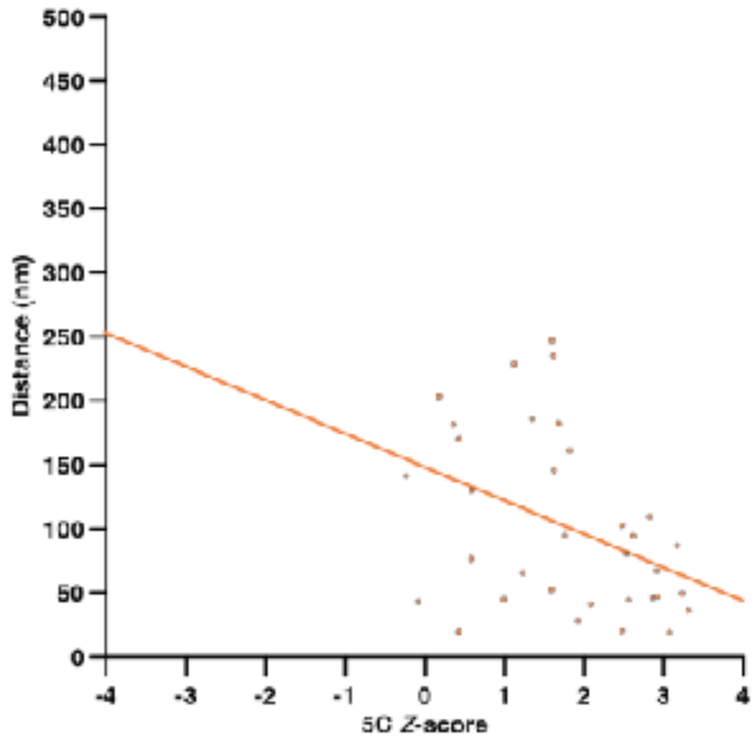
Neighbor fragments

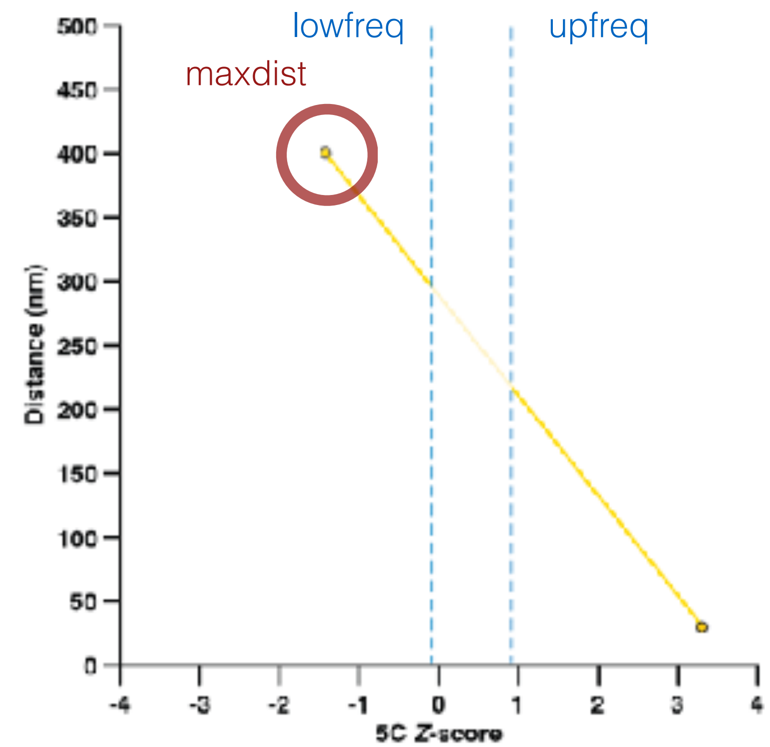


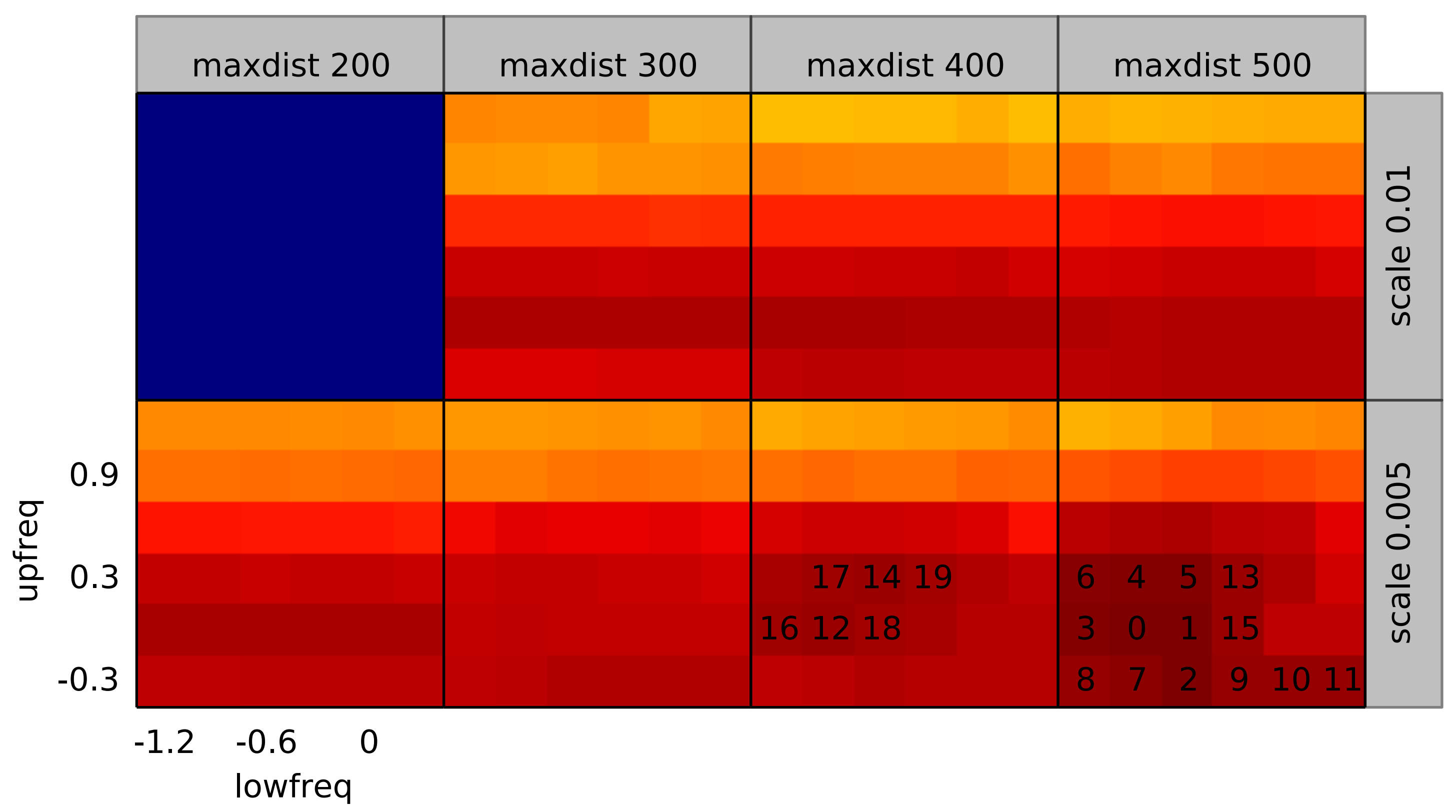
Non-Neighbor fragments



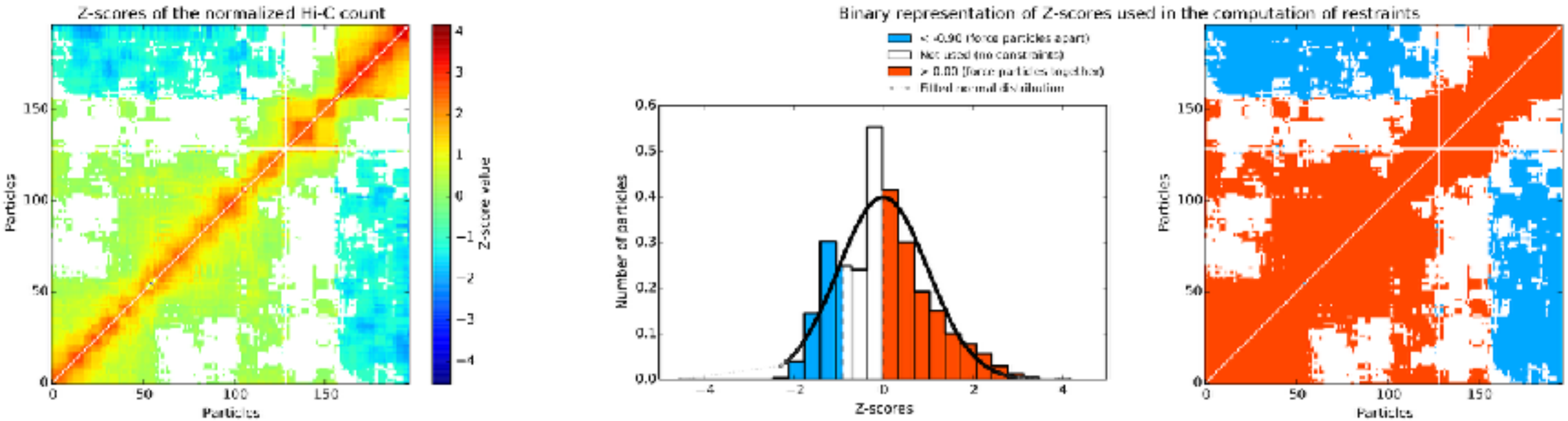
Parameter optimization

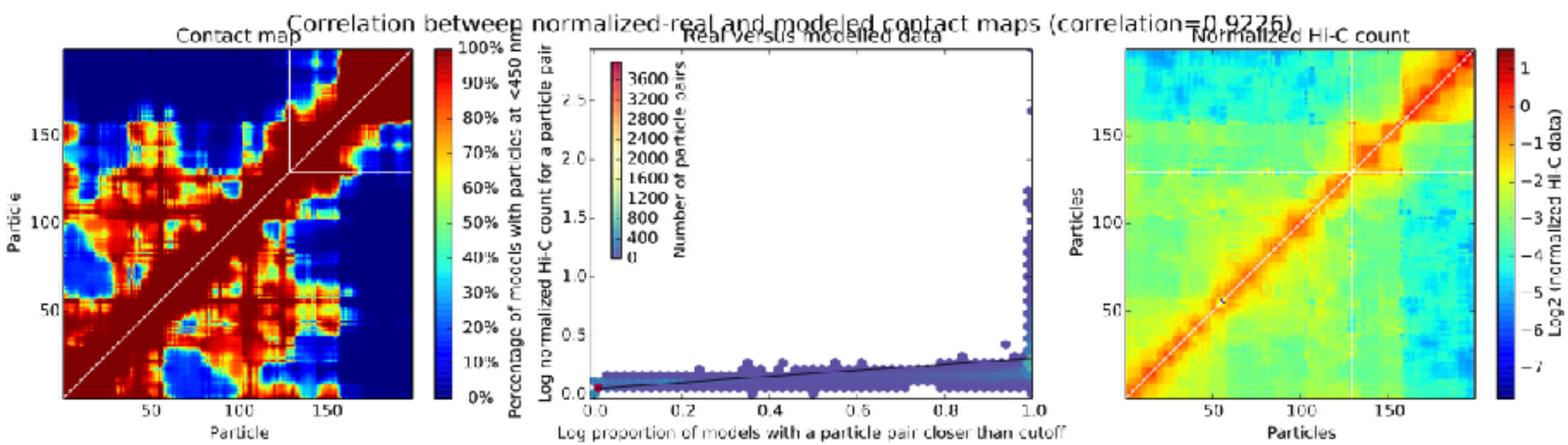






Parameter optimization



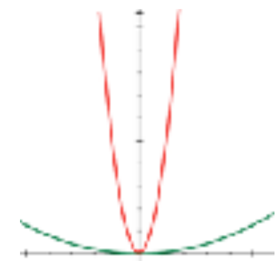


Representation

Constituent parts of the molecule

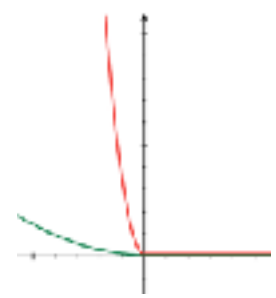
Harmonic

$$H_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2$$



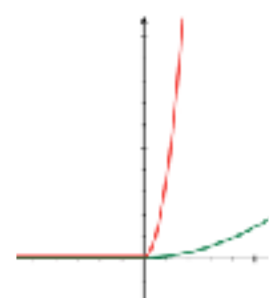
Harmonic Lower Bound

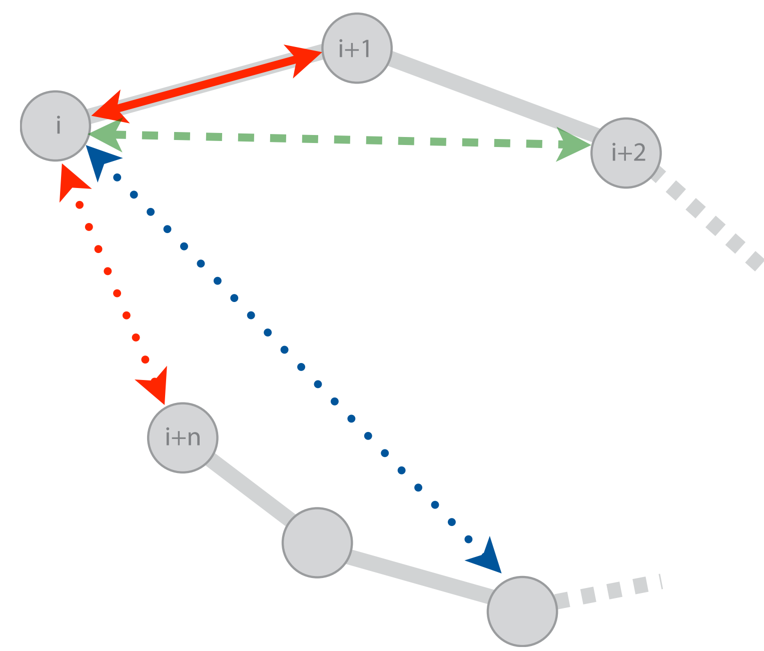
$$\begin{cases} if\ \ d_{i,j} \le d^0_{i,j}; & lbH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} > d^0_{i,j}; & lbH_{i,j} = 0 \end{cases}$$



Harmonic Upper Bound

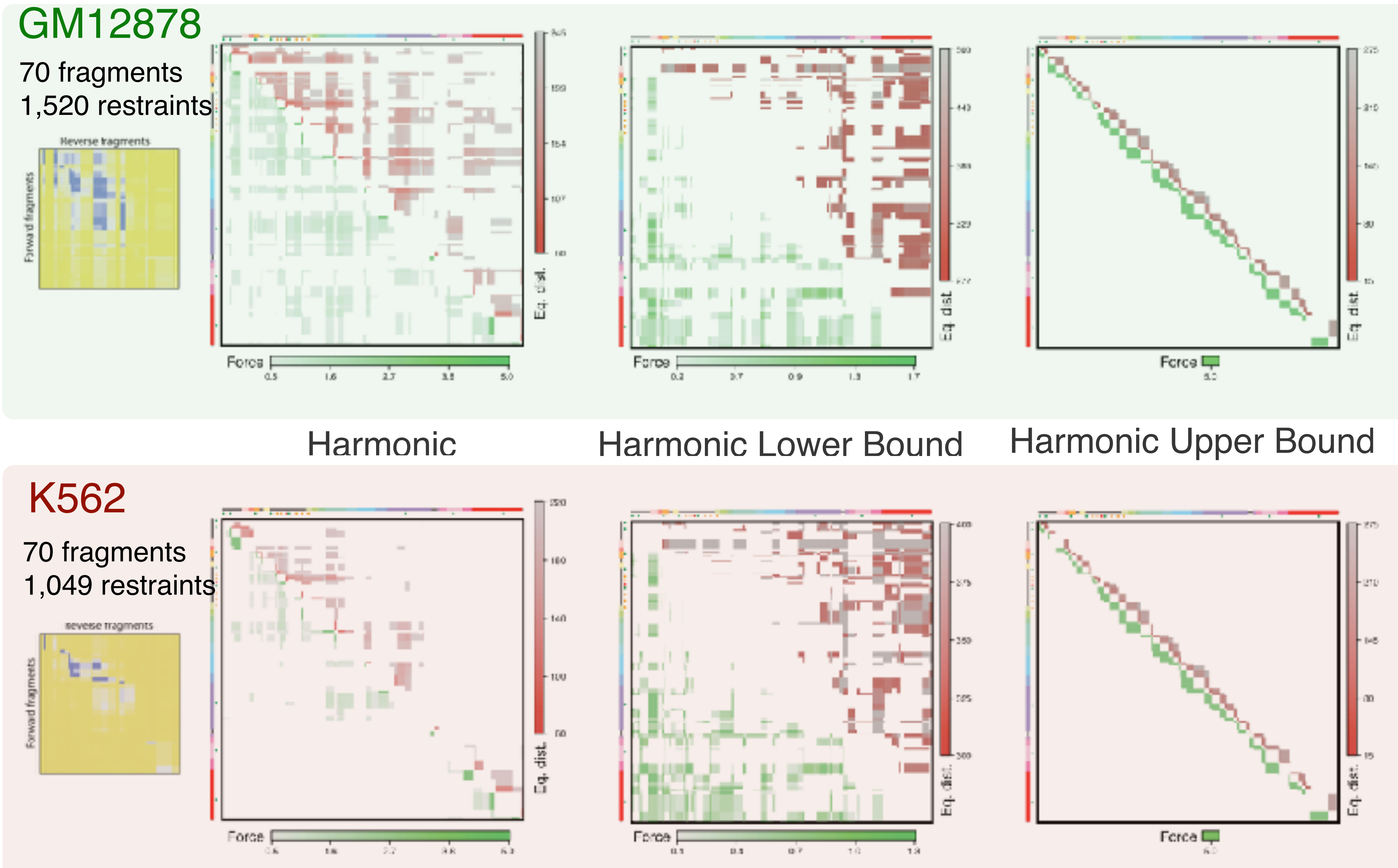
$$\begin{cases} if\ \ d_{i,j} \ge d^0_{i,j}; & ubH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} < d^0_{i,j}; & ubH_{i,j} = 0 \end{cases}$$



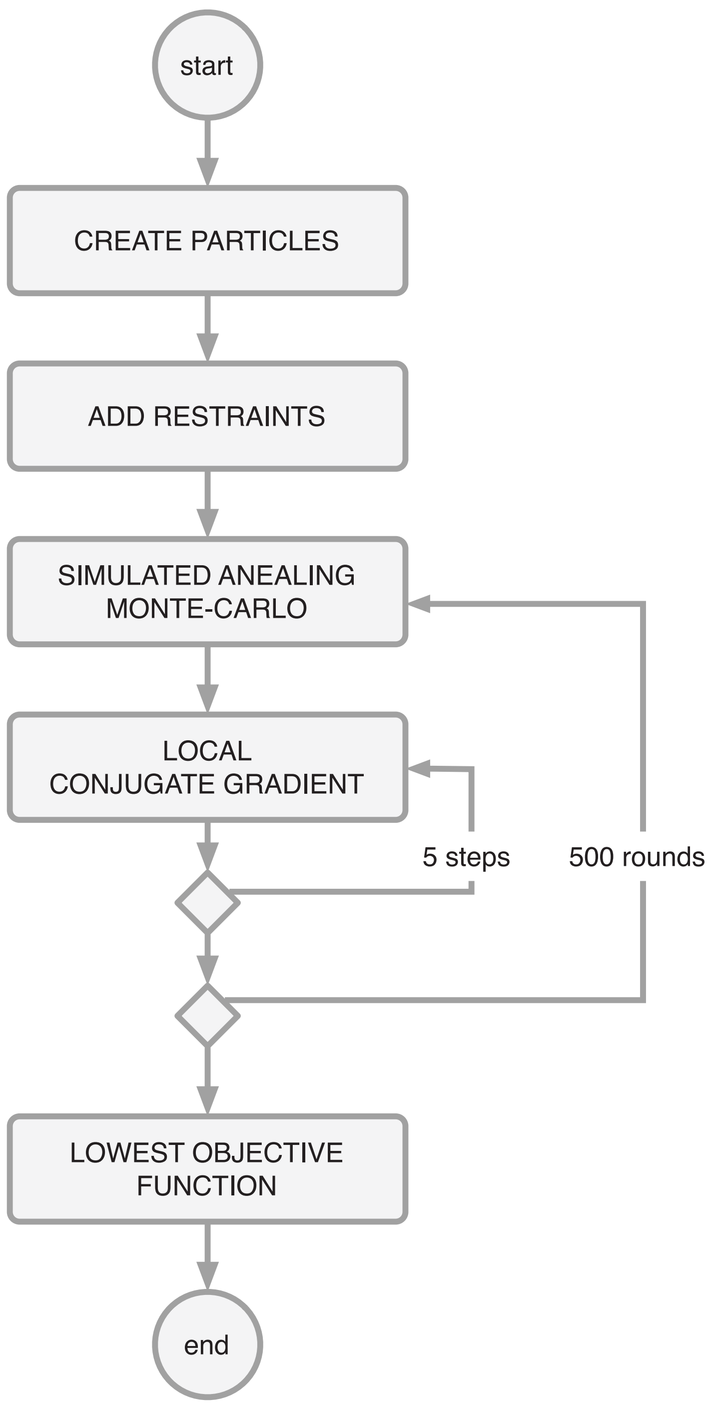


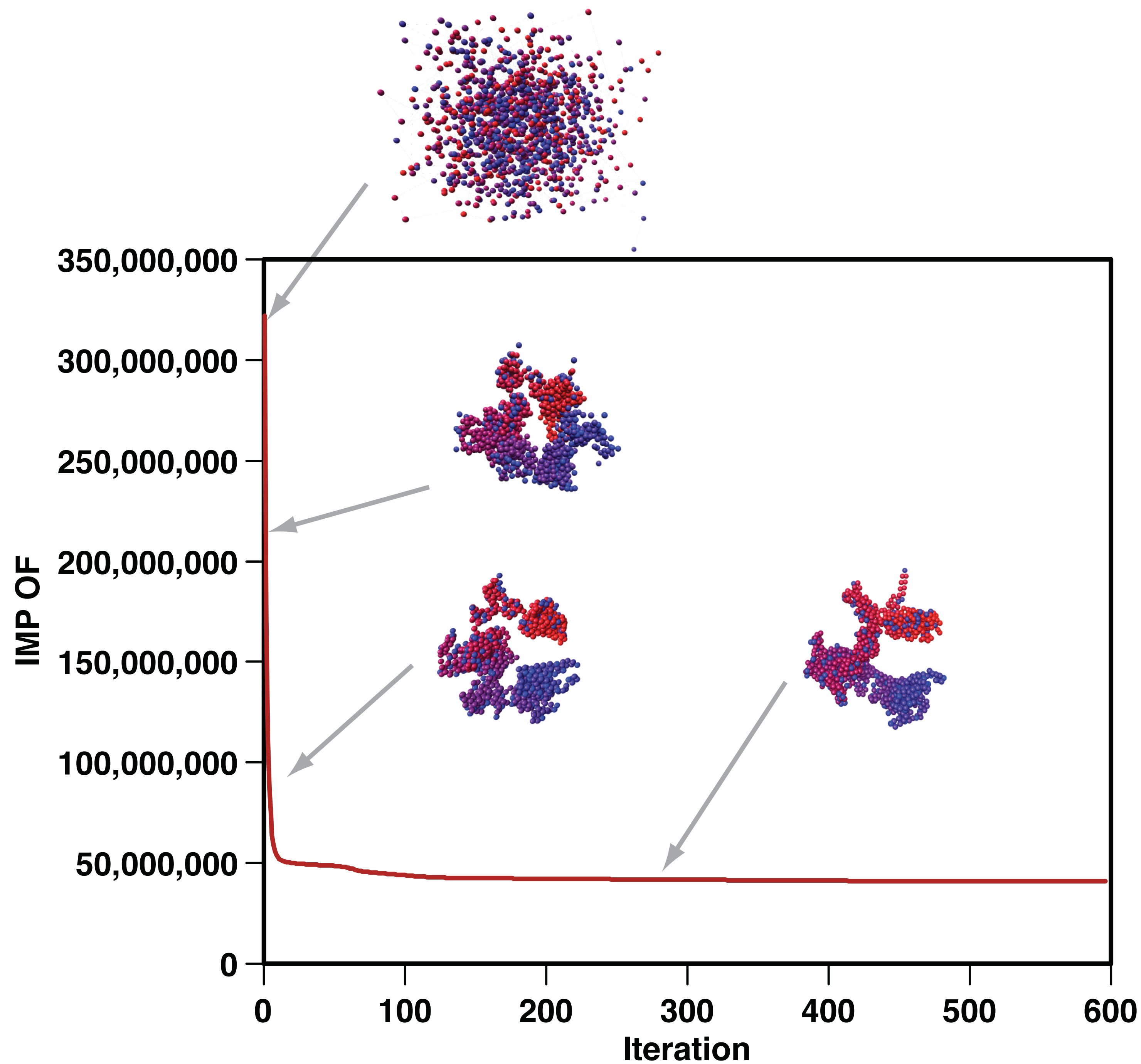
Scoring

Individual spatial restraints encoding the data



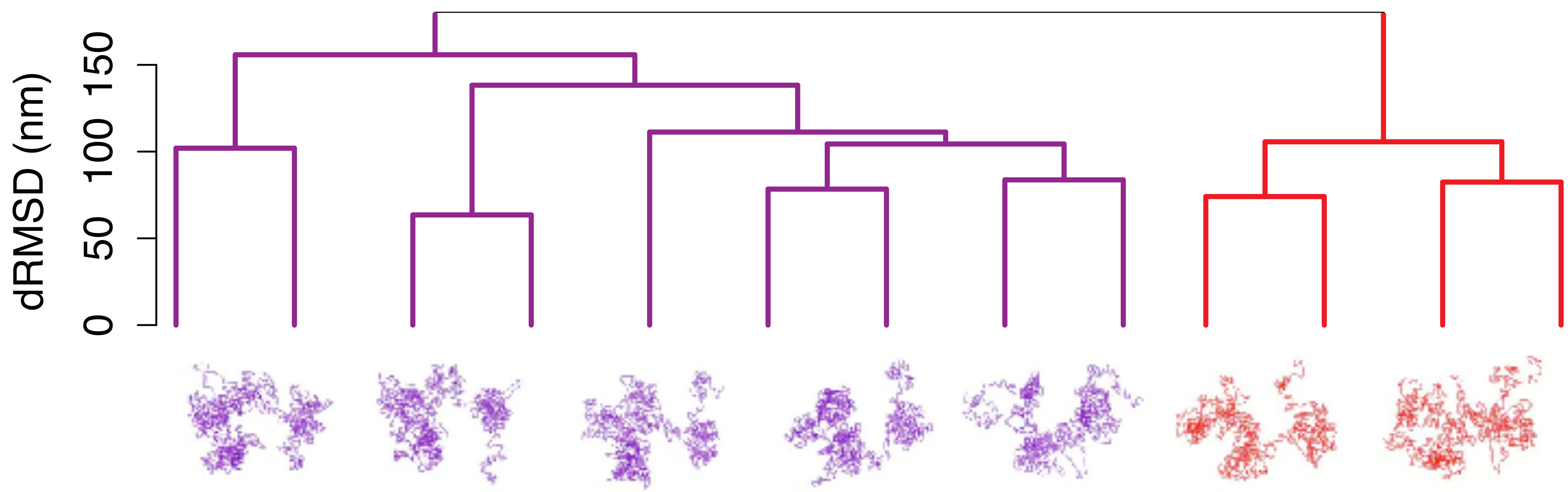
Optimization of the scoring function

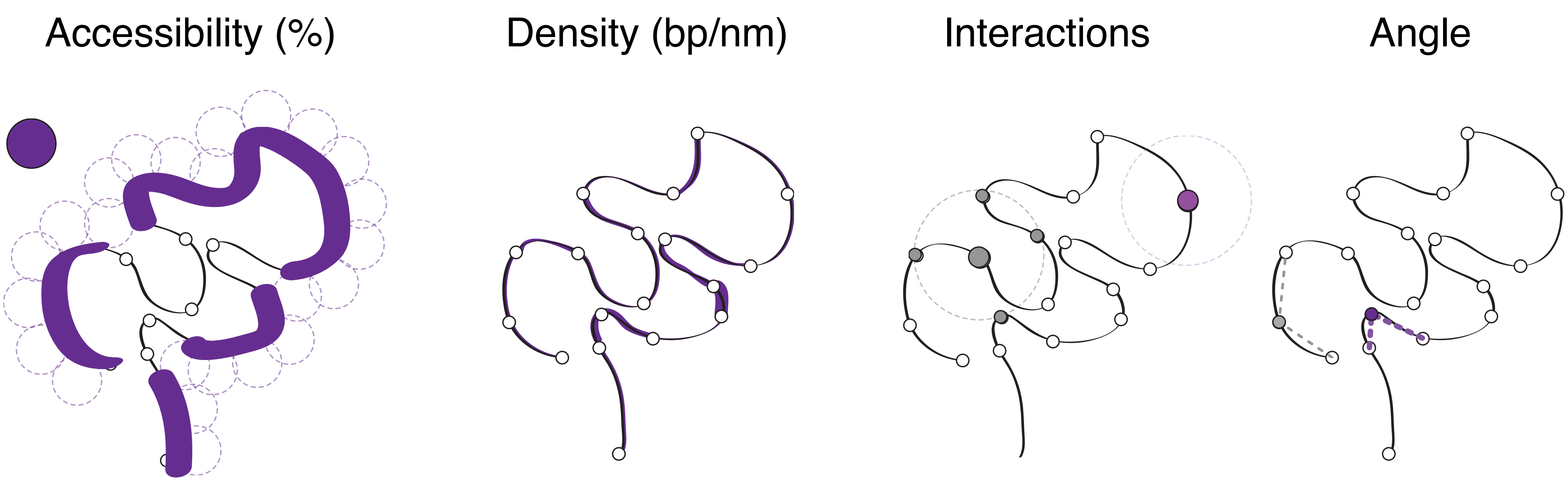




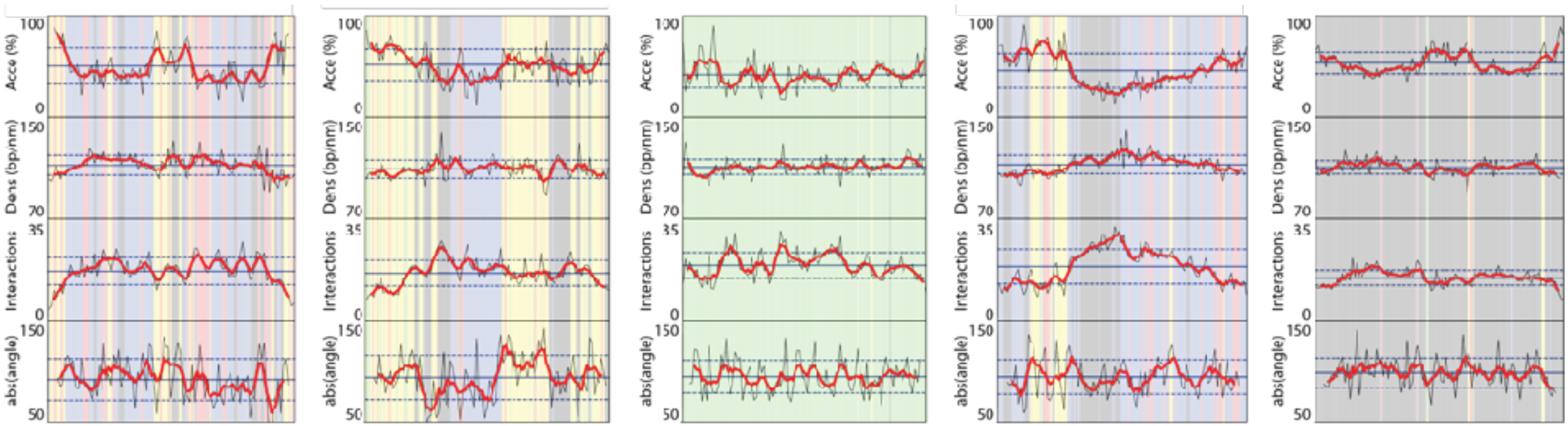
cnag CRG

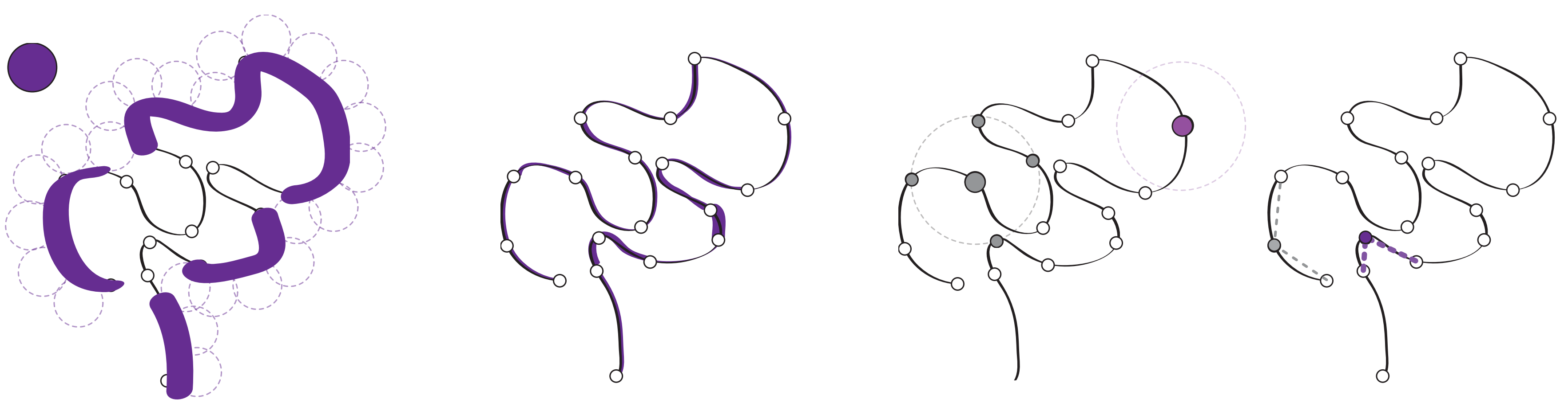
Model analysis: clustering and structural features





Model analysis: Structural features

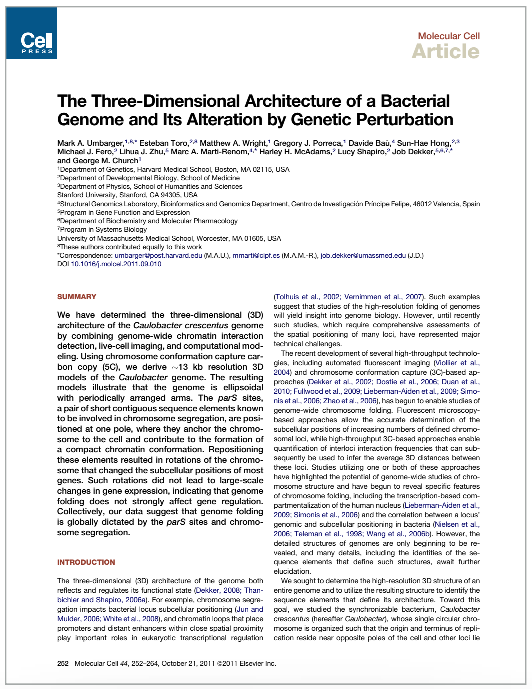


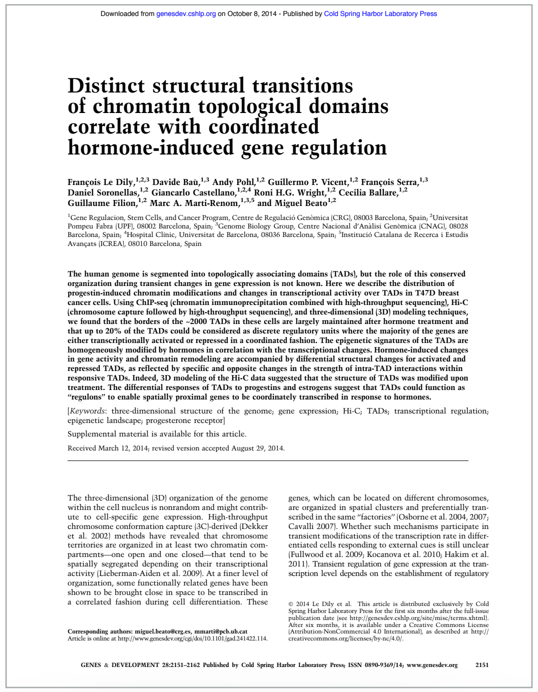


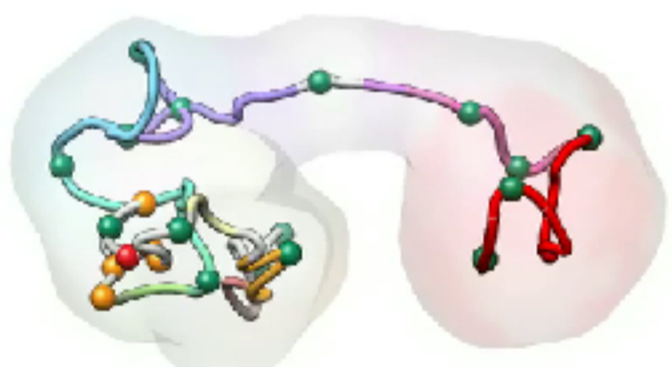
TADbit previous applications...

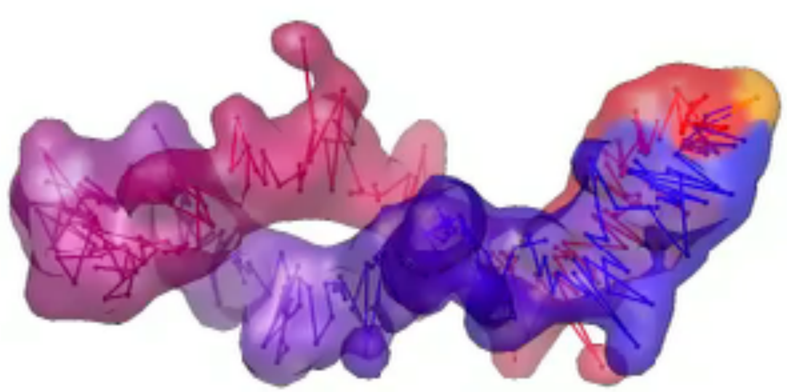
Baù, D. et al. Nat Struct Mol Biol (2011)
Umbarger, M. A. et al. Mol Cell (2011)
Le Dily, F. et al. Genes & Dev (2014)

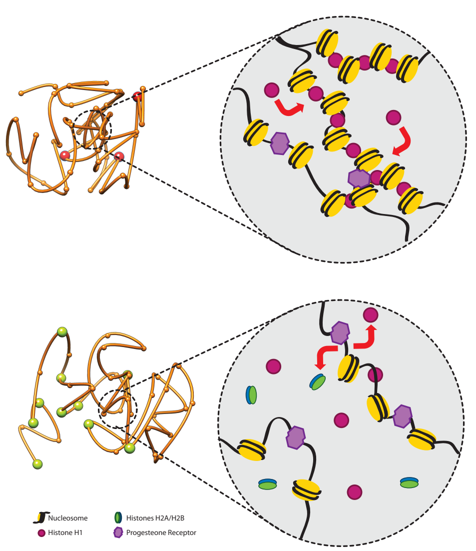

TECHNICAL REPORTS
nature structural & molecular biology
The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules
Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker & Marc A Marti-Renom


Cell PRESS
Molecular Cell Article
The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation

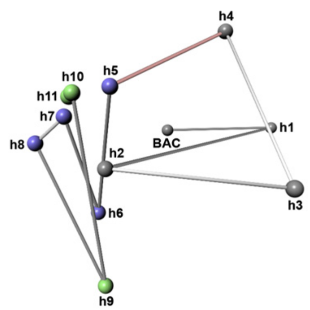

Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation



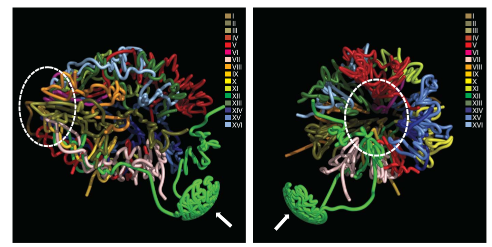




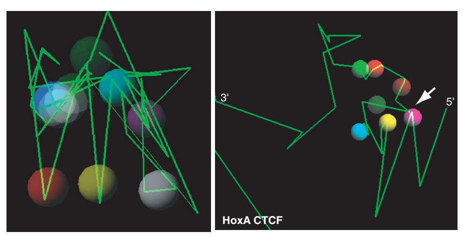
Are the models correct?



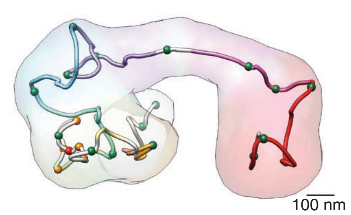
Jhunjhunwala (2008) Cell



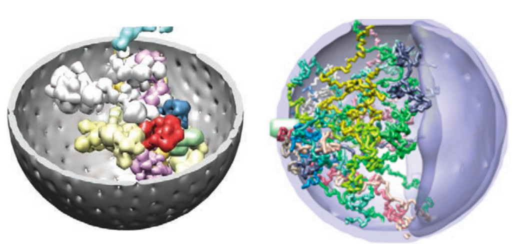
Duan (2010) Nature



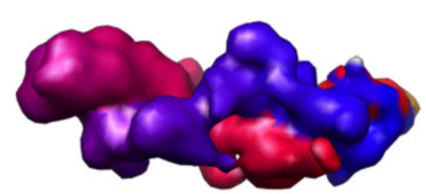
Fraser (2009) Genome Biology
Ferraiuolo (2010) Nucleic Acids Research



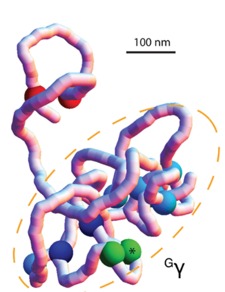
Baù (2011) Nature Structural & Molecular Biology



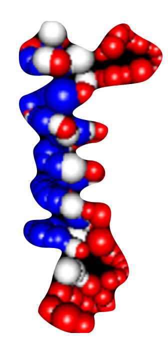
Kalhor (2011) Nature Biotechnology
Tjong (2012) Genome Research



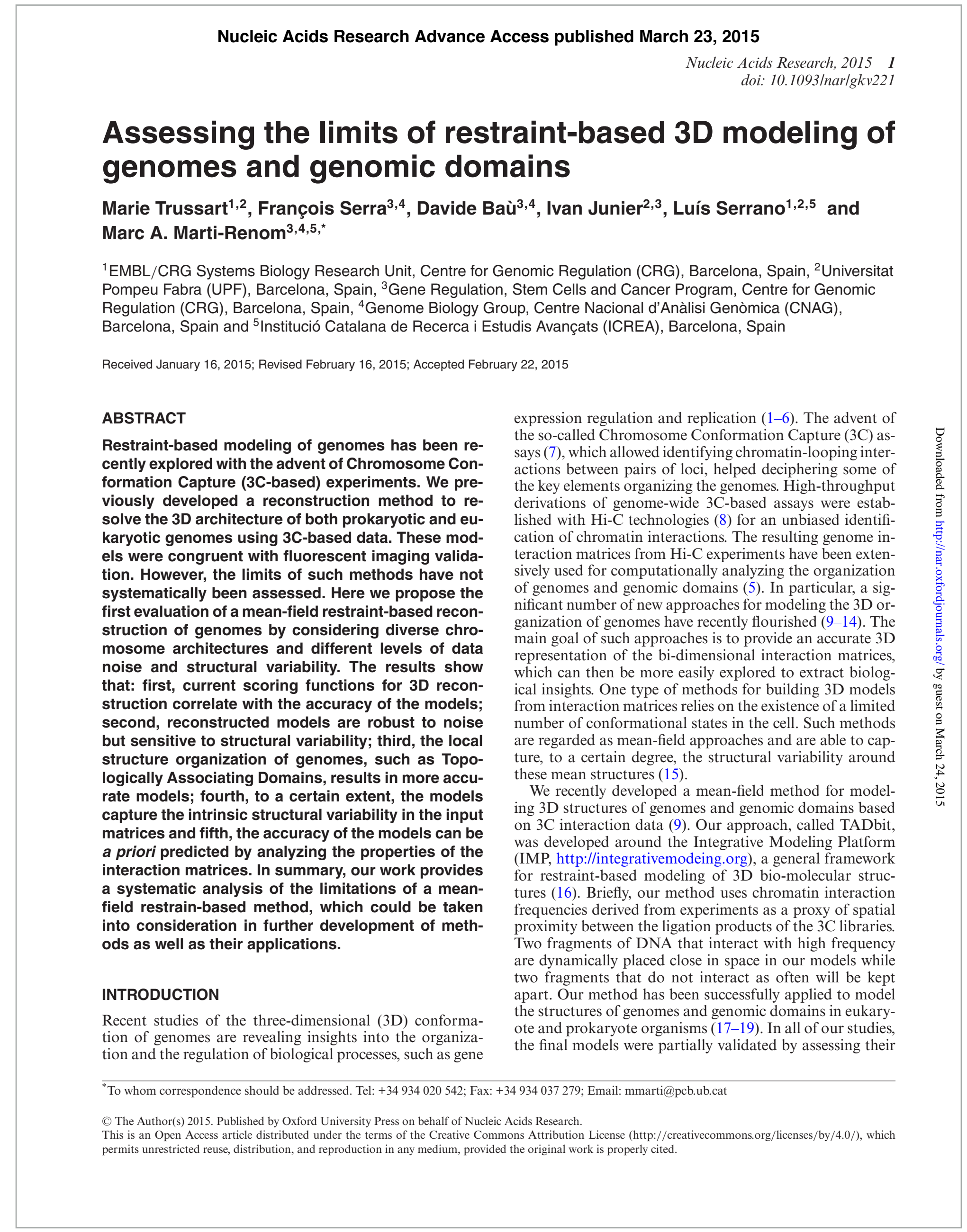
Umbarger (2011) Molecular Cell



Junier (2012) Nucleic Acids Research



Hu (2013) PLoS Computational Biology



Nucleic Acids Research Advance Access published March 23, 2015

Nucleic Acids Research, 2015 1
doi: 10.1093/nar/gkv221

Assessing the limits of restraint-based 3D modeling of genomes and genomic domains

Marie Trussart[1,2], François Serra[3,4], Davide Baù[3,4], Ivan Junier[2,3], Luís Serrano[1,2,5] and Marc A. Marti-Renom[3,4,5,*]

[1]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, [4]Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain and [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Received January 16, 2015; Revised February 16, 2015; Accepted February 22, 2015

ABSTRACT

Restraint-based modeling of genomes has been recently explored with the advent of Chromosome Conformation Capture (3C-based) experiments. We previously developed a reconstruction method to resolve the 3D architecture of both prokaryotic and eukaryotic genomes using 3C-based data. These models were congruent with fluorescent imaging validation. However, the limits of such methods have not systematically been assessed. Here we propose the first evaluation of a mean-field restraint-based reconstruction of genomes by considering diverse chromosome architectures and different levels of data noise and structural variability. The results show that: first, current scoring functions for 3D reconstruction correlate with the accuracy of the models; second, reconstructed models are robust to noise but sensitive to structural variability; third, the local structure organization of genomes, such as Topologically Associating Domains, results in more accurate models; fourth, to a certain extent, the models capture the intrinsic structural variability in the input matrices and fifth, the accuracy of the models can be *a priori* predicted by analyzing the properties of the interaction matrices. In summary, our work provides a systematic analysis of the limitations of a mean-field restrain-based method, which could be taken into consideration in further development of methods as well as their applications.

INTRODUCTION

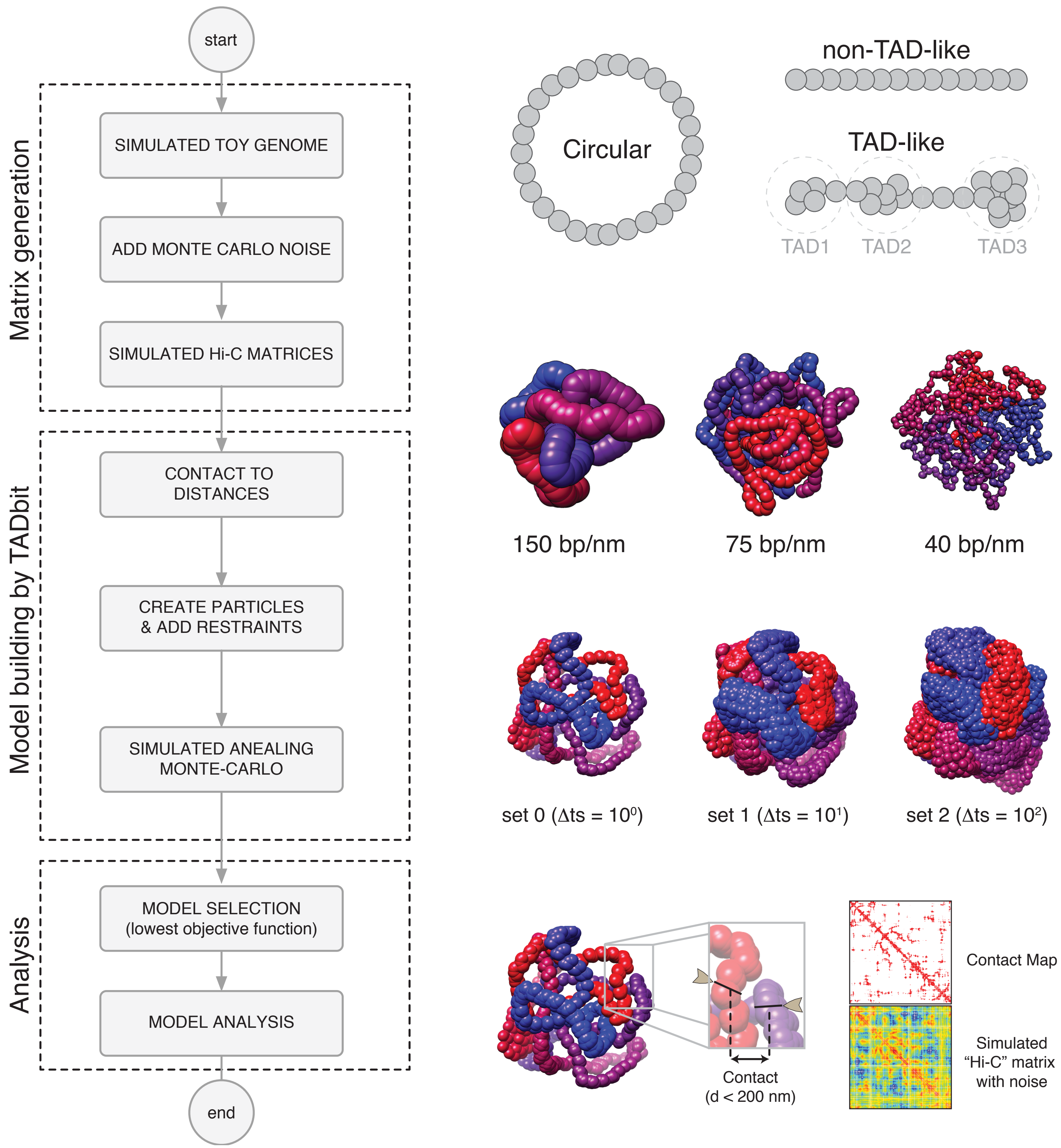
Recent studies of the three-dimensional (3D) conformation of genomes are revealing insights into the organization and the regulation of biological processes, such as gene expression regulation and replication (1–6). The advent of the so-called Chromosome Conformation Capture (3C) assays (7), which allowed identifying chromatin-looping interactions between pairs of loci, helped deciphering some of the key elements organizing the genomes. High-throughput derivations of genome-wide 3C-based assays were established with Hi-C technologies (8) for an unbiased identification of chromatin interactions. The resulting genome interaction matrices from Hi-C experiments have been extensively used for computationally analyzing the organization of genomes and genomic domains (5). In particular, a significant number of new approaches for modeling the 3D organization of genomes have recently flourished (9–14). The main goal of such approaches is to provide an accurate 3D representation of the bi-dimensional interaction matrices, which can then be more easily explored to extract biological insights. One type of methods for building 3D models from interaction matrices relies on the existence of a limited number of conformational states in the cell. Such methods are regarded as mean-field approaches and are able to capture, to a certain degree, the structural variability around these mean structures (15).

We recently developed a mean-field method for modeling 3D structures of genomes and genomic domains based on 3C interaction data (9). Our approach, called TADbit, was developed around the Integrative Modeling Platform (IMP, http://integrativemodeing.org), a general framework for restraint-based modeling of 3D bio-molecular structures (16). Briefly, our method uses chromatin interaction frequencies derived from experiments as a proxy of spatial proximity between the ligation products of the 3C libraries. Two fragments of DNA that interact with high frequency are dynamically placed close in space in our models while two fragments that do not interact as often will be kept apart. Our method has been successfully applied to model the structures of genomes and genomic domains in eukaryote and prokaryote organisms (17–19). In all of our studies, the final models were partially validated by assessing their

*To whom correspondence should be addressed. Tel: +34 934 020 542; Fax: +34 934 037 279; Email: mmarti@pcb.ub.cat

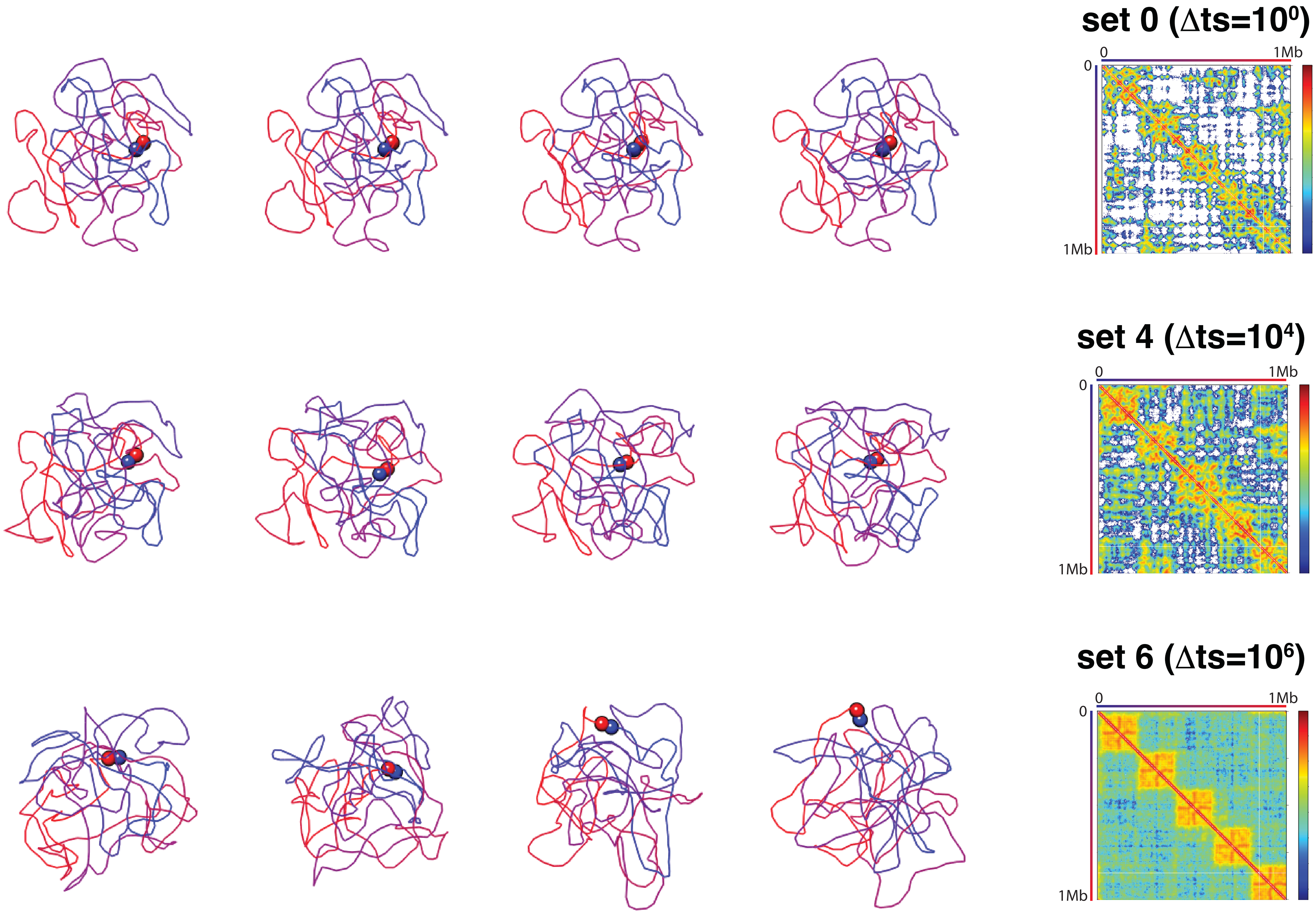
Trussart, et al. (2015). Nucleic Acids Research.

Toy models

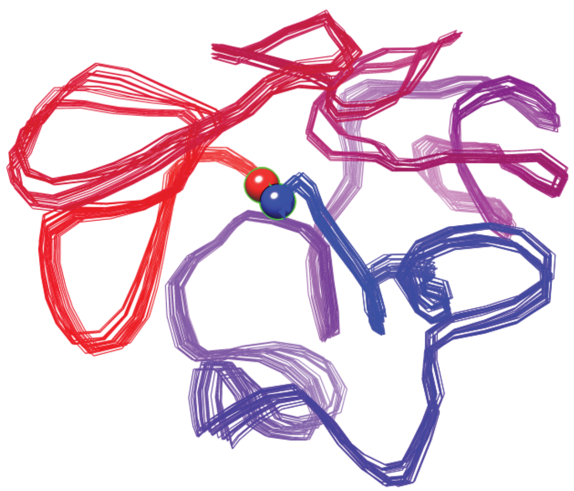


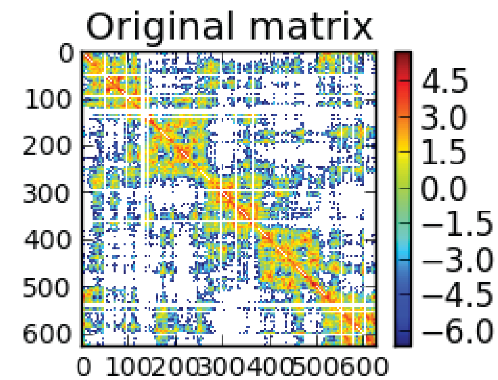
by Ivan Junier

Toy interaction matrices

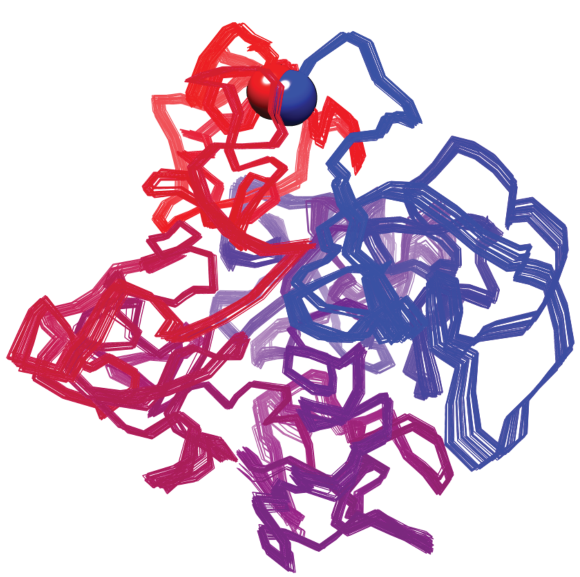


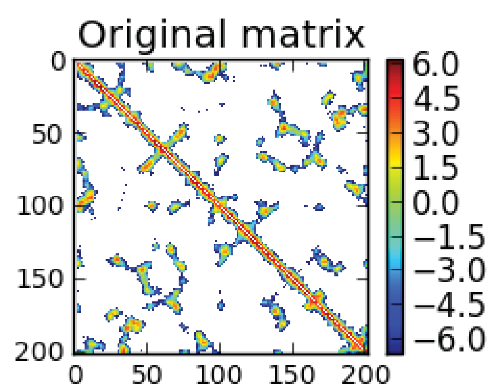
Reconstructing toy models

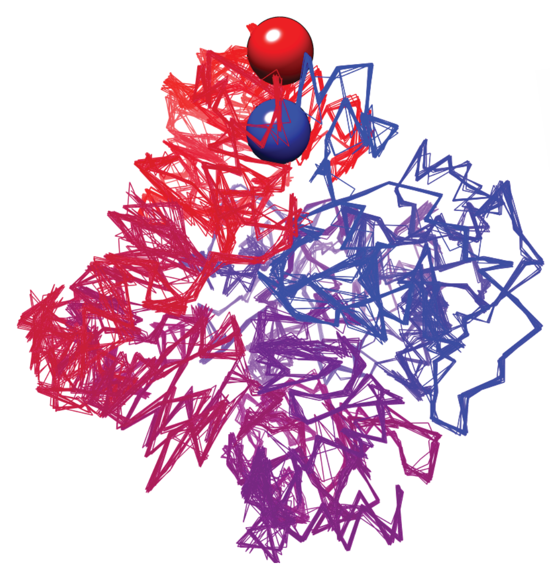




chr40_TAD
α=100
Δts=10
TADbit-SCC: 0.91
<dRMSD>: 32.7 nm
<dSCC>: 0.94

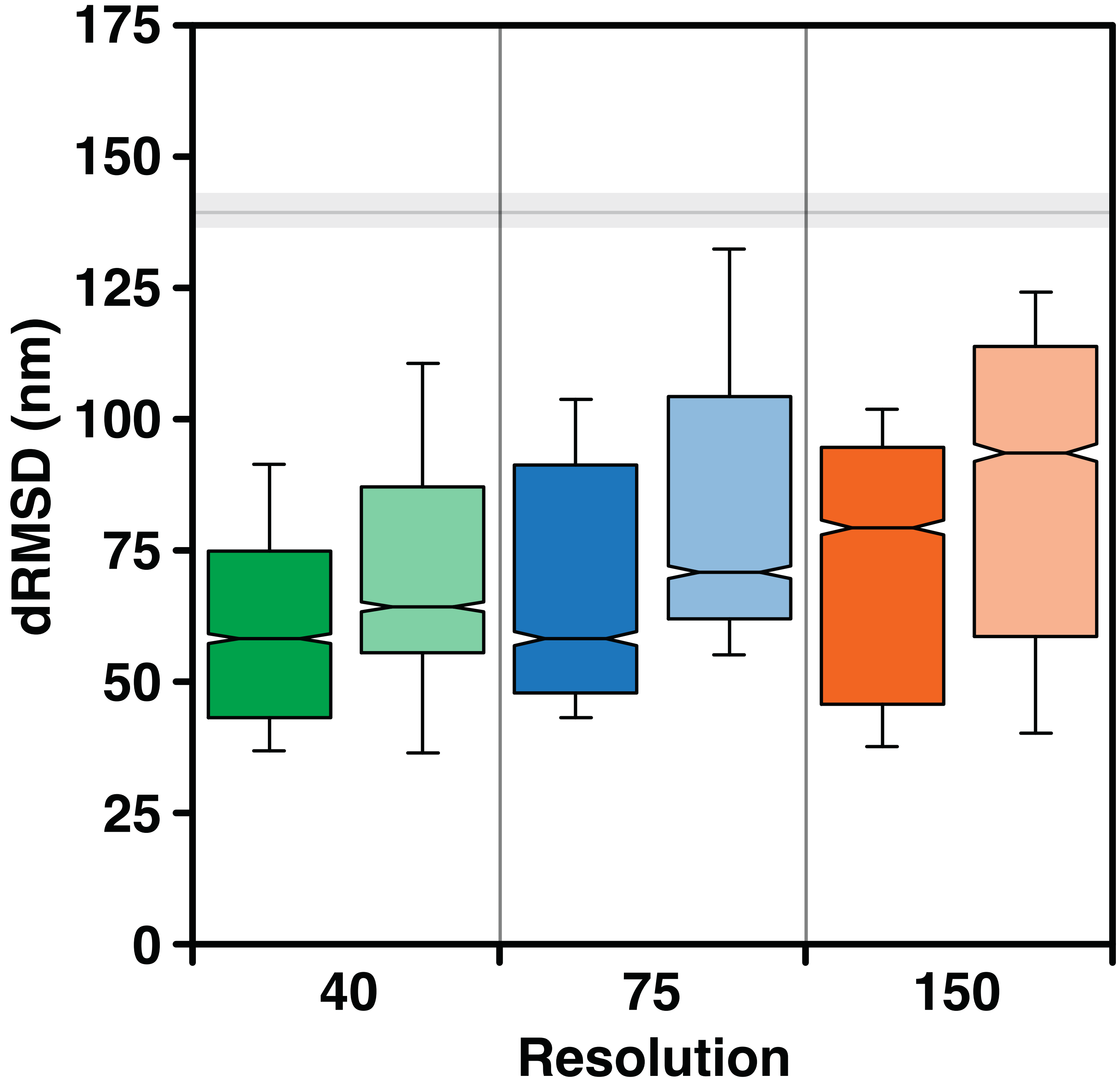




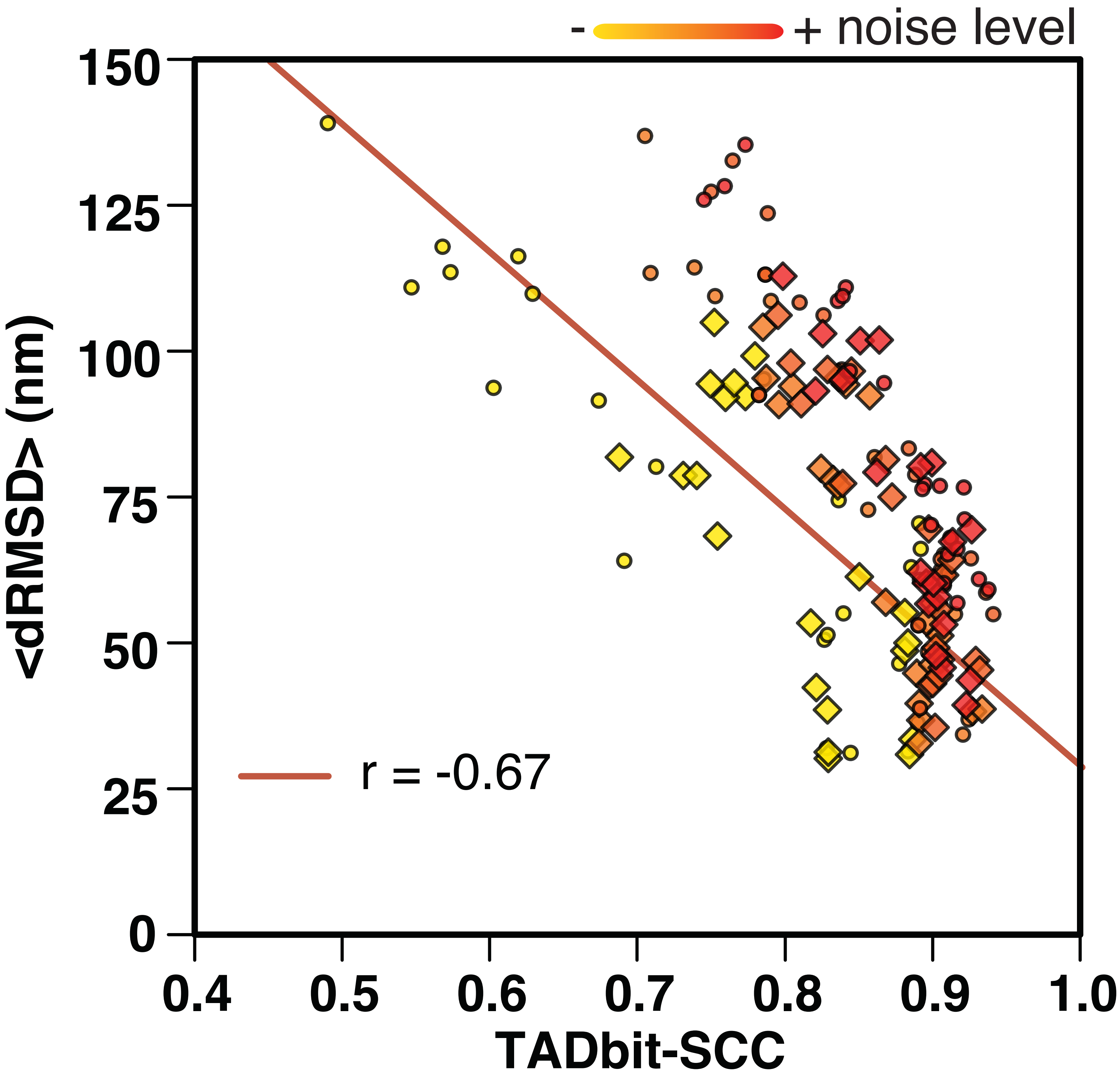


chr150_TAD
α=50
Δts=1
TADbit-SCC: 0.82
<dRMSD>: 45.4 nm
<dSCC>: 0.86

TADs & higher-res are "good"

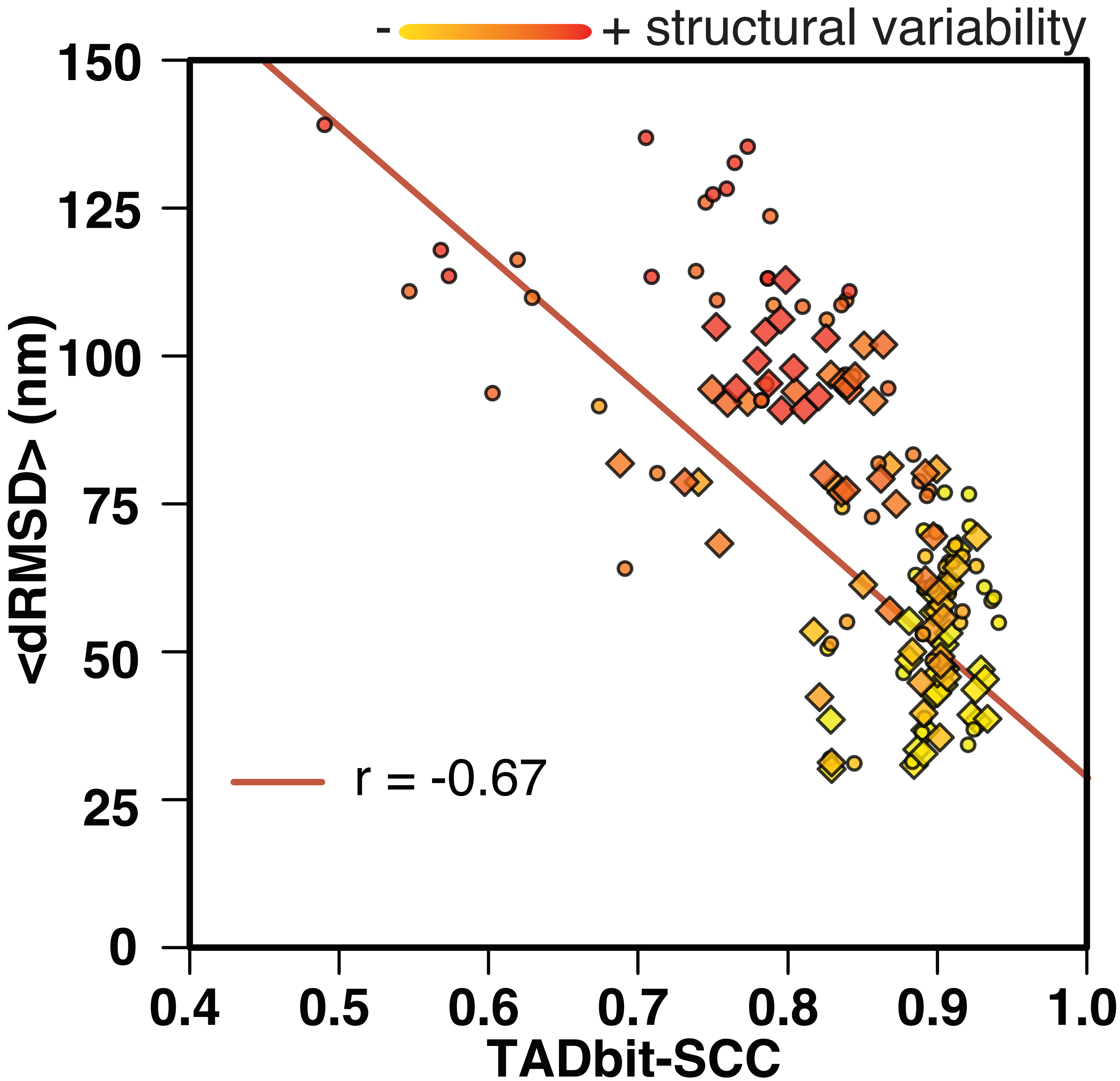


Noise is "OK"

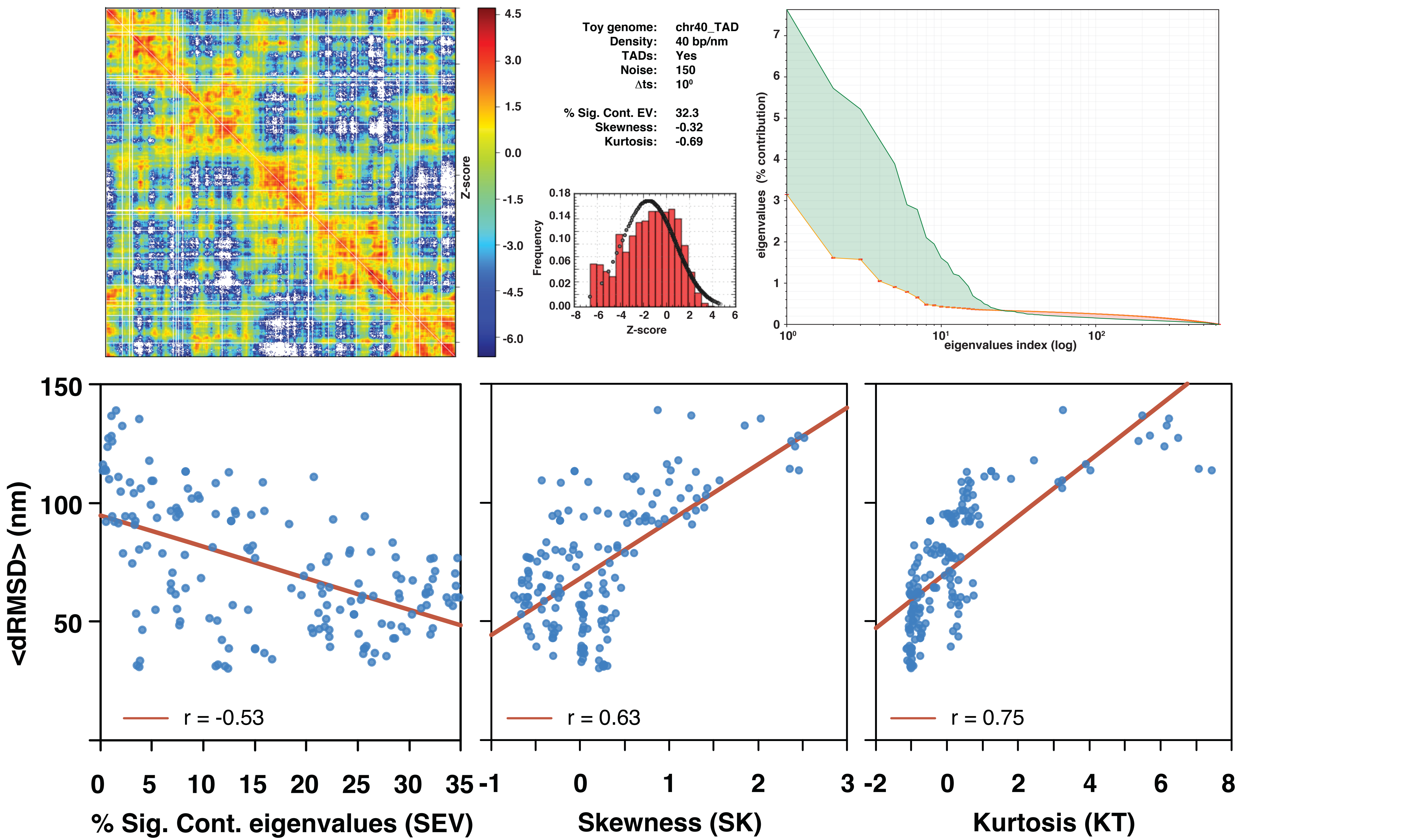


- + noise level
150
125
100
75
50
25
0
<dRMSD> (nm)
0.4
0.5
0.6
0.7
0.8
0.9
1.0
TADbit-SCC
r = -0.67

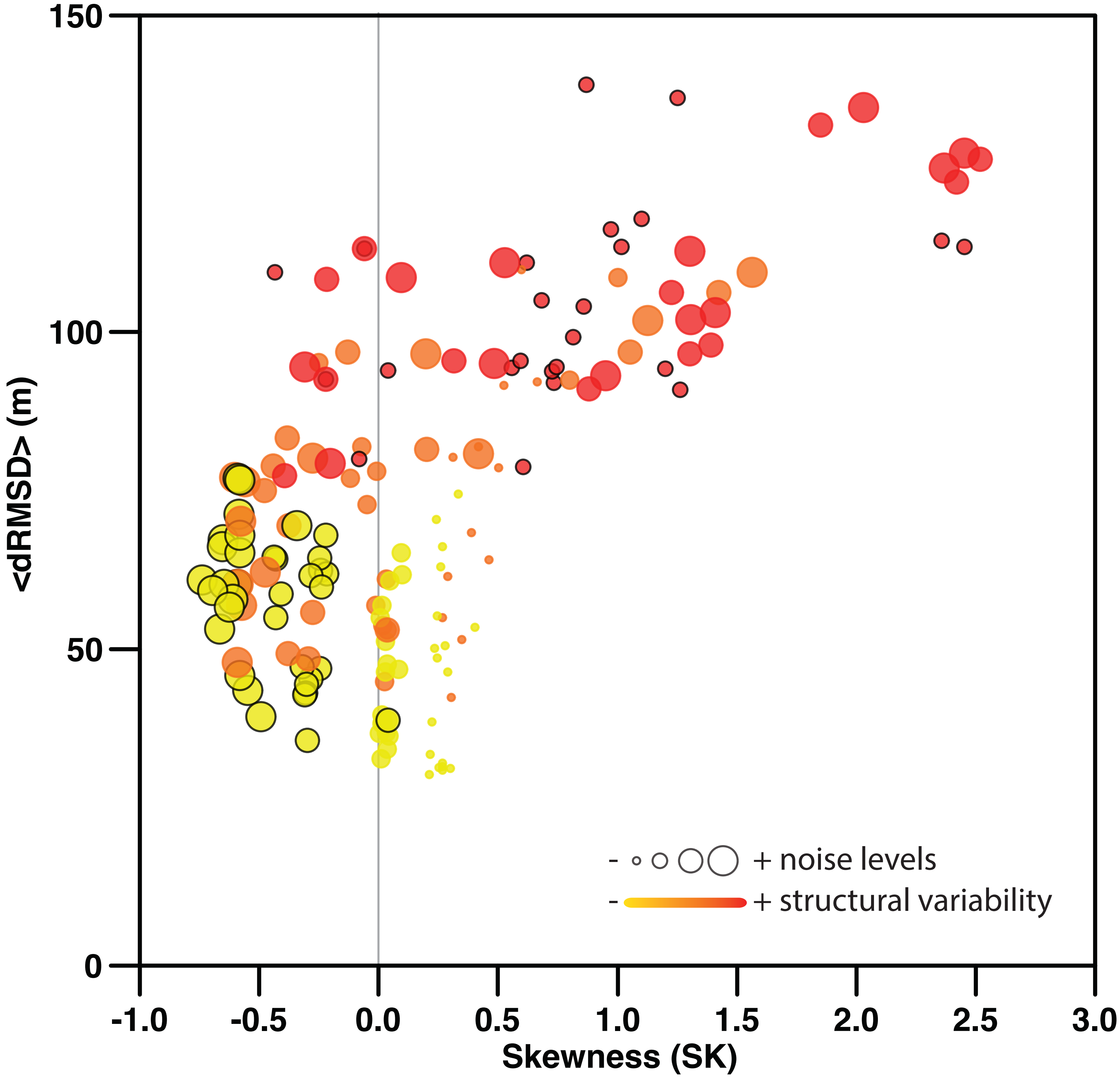
Structural variability is “NOT OK”



Can we predict the accuracy of the models?

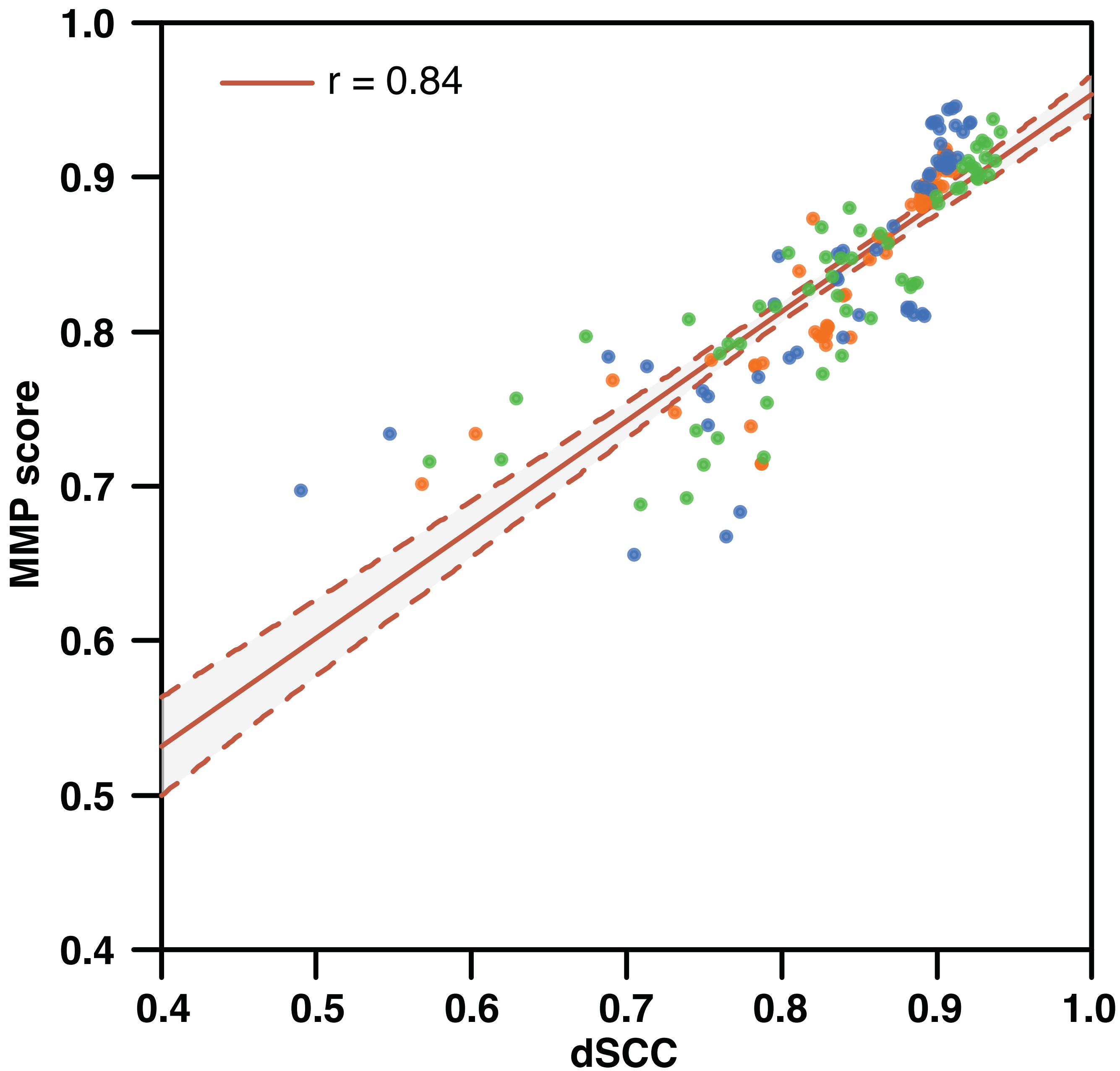


Skewness "side effect"

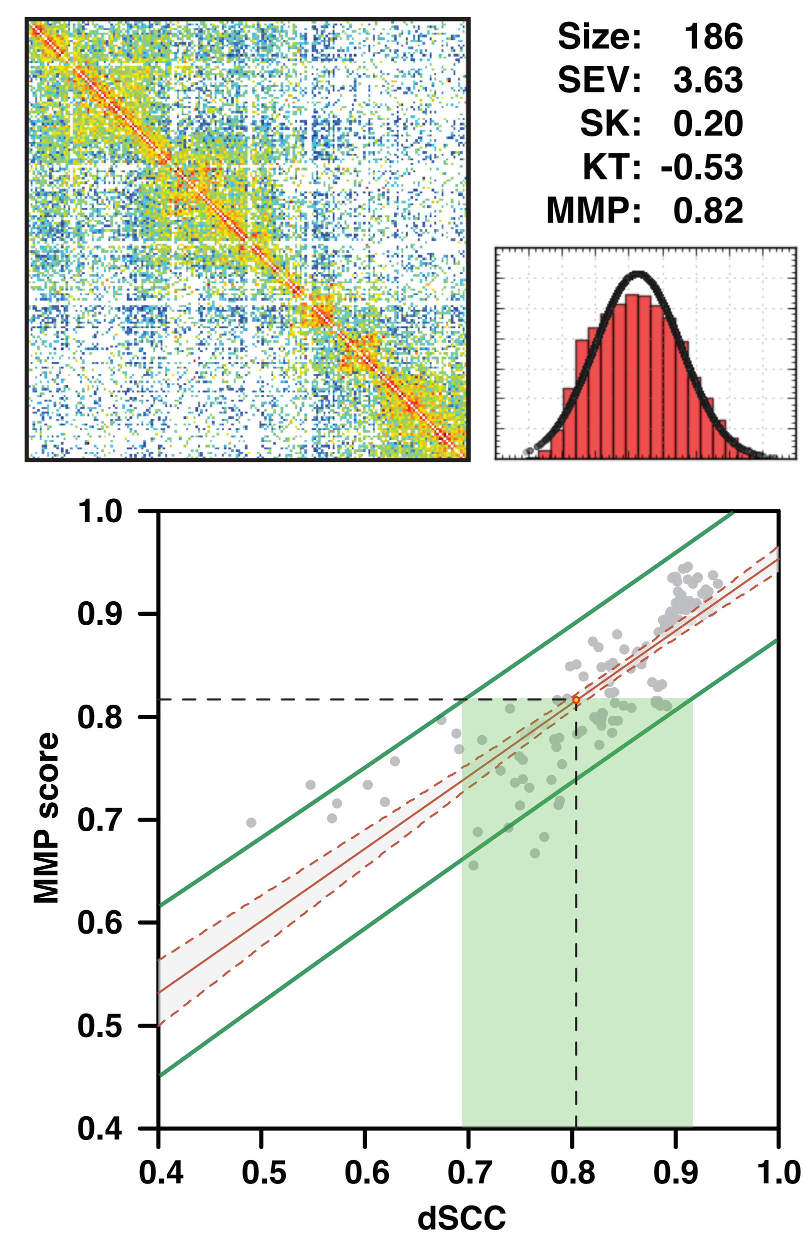


Can we predict the accuracy of the models?

$$MMP = -0.0002 * \text{Size} + 0.0335 * \text{SK} - 0.0229 * \text{KU} + 0.0069 * \text{SEV} + 0.8126$$



Human Chr1:120,640,000-128,040,000



Higher-res is “good”

put your $$ in sequencing

Noise is “OK”

no need to worry much

Structural variability is ”NOT OK”

homogenize your cell population!

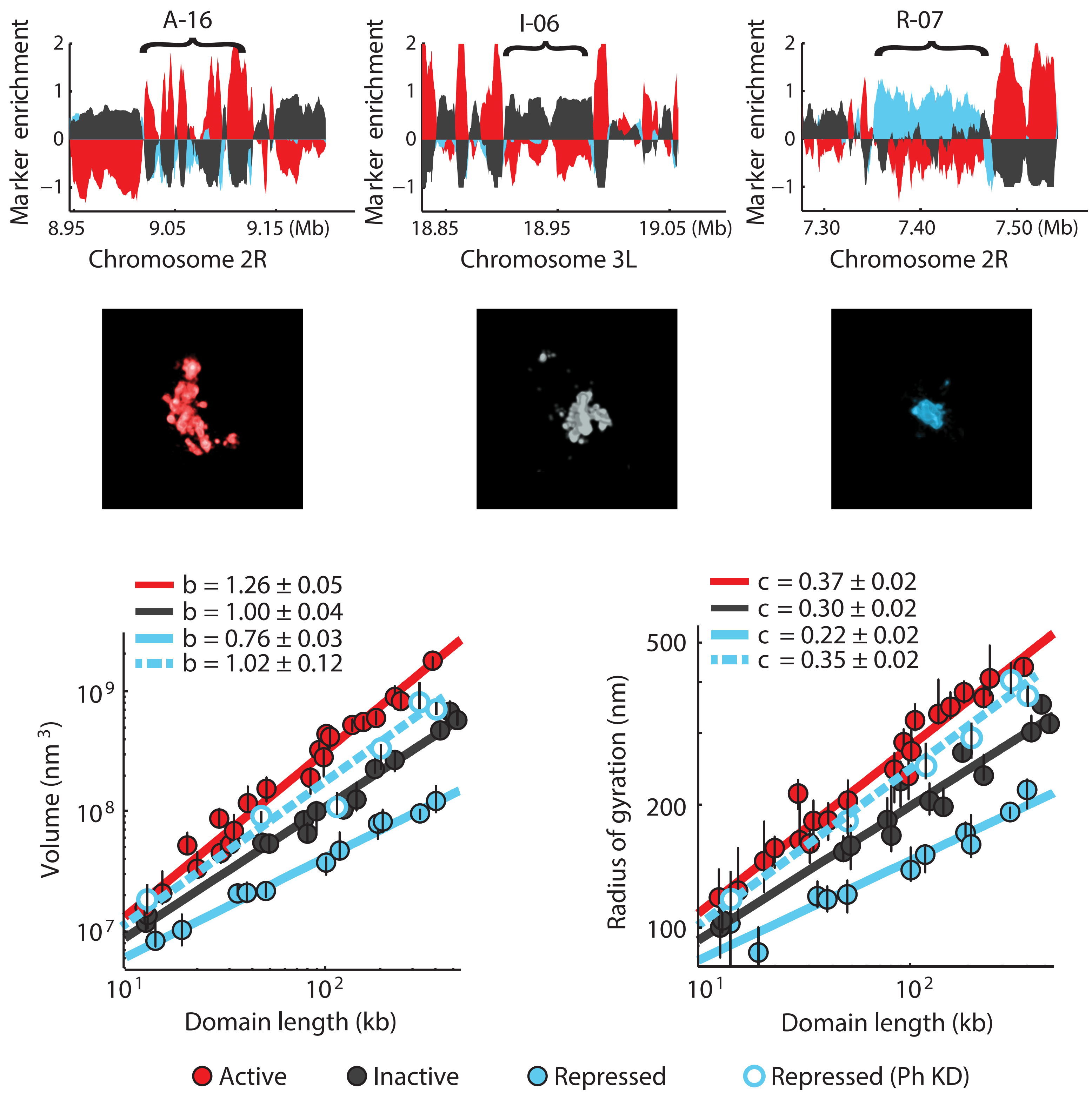
…but we can differentiate between noise and structural variability

and we can a priori predict the accuracy of the models

But… what about direct validation of models?

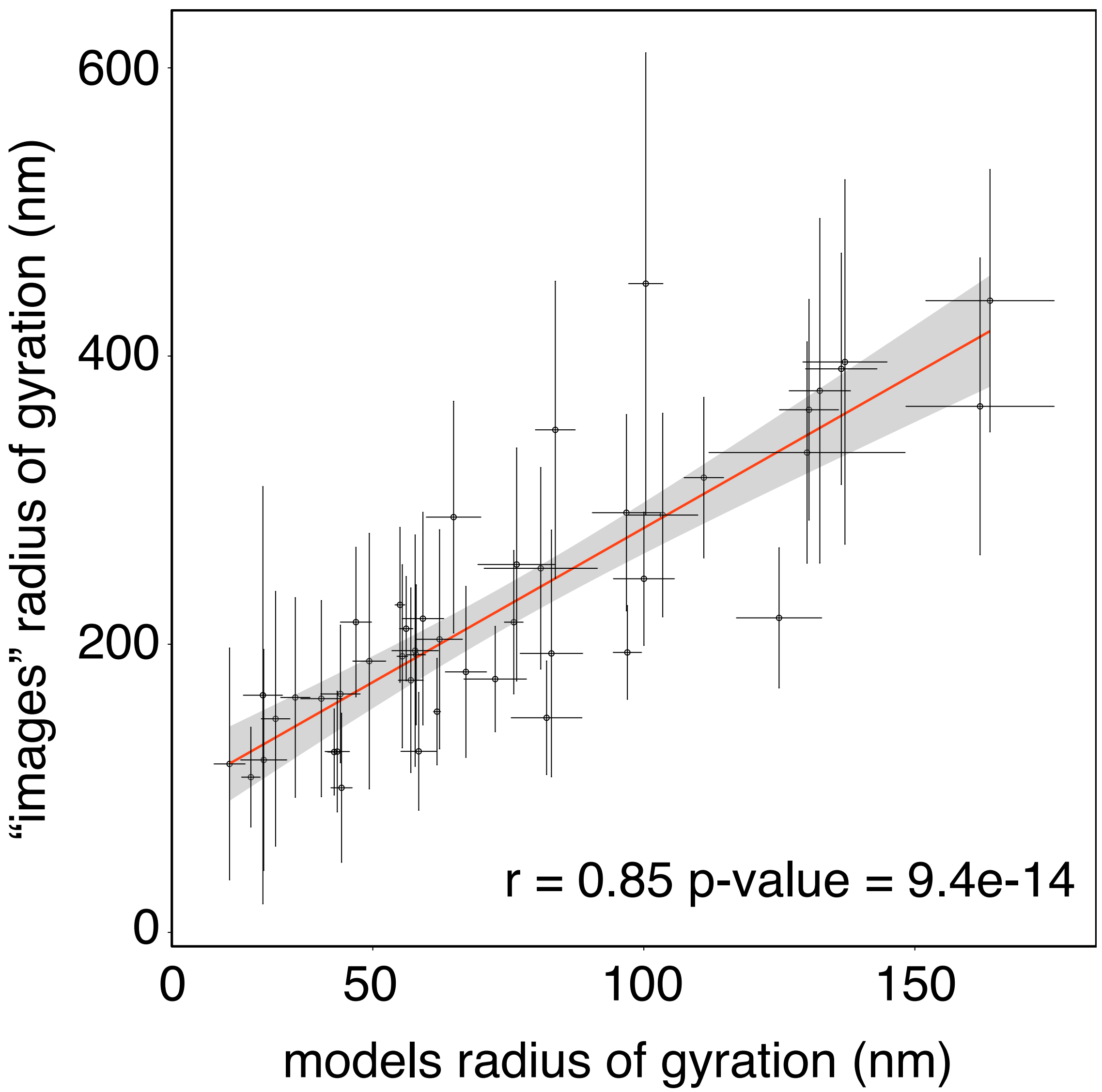
Model accuracy

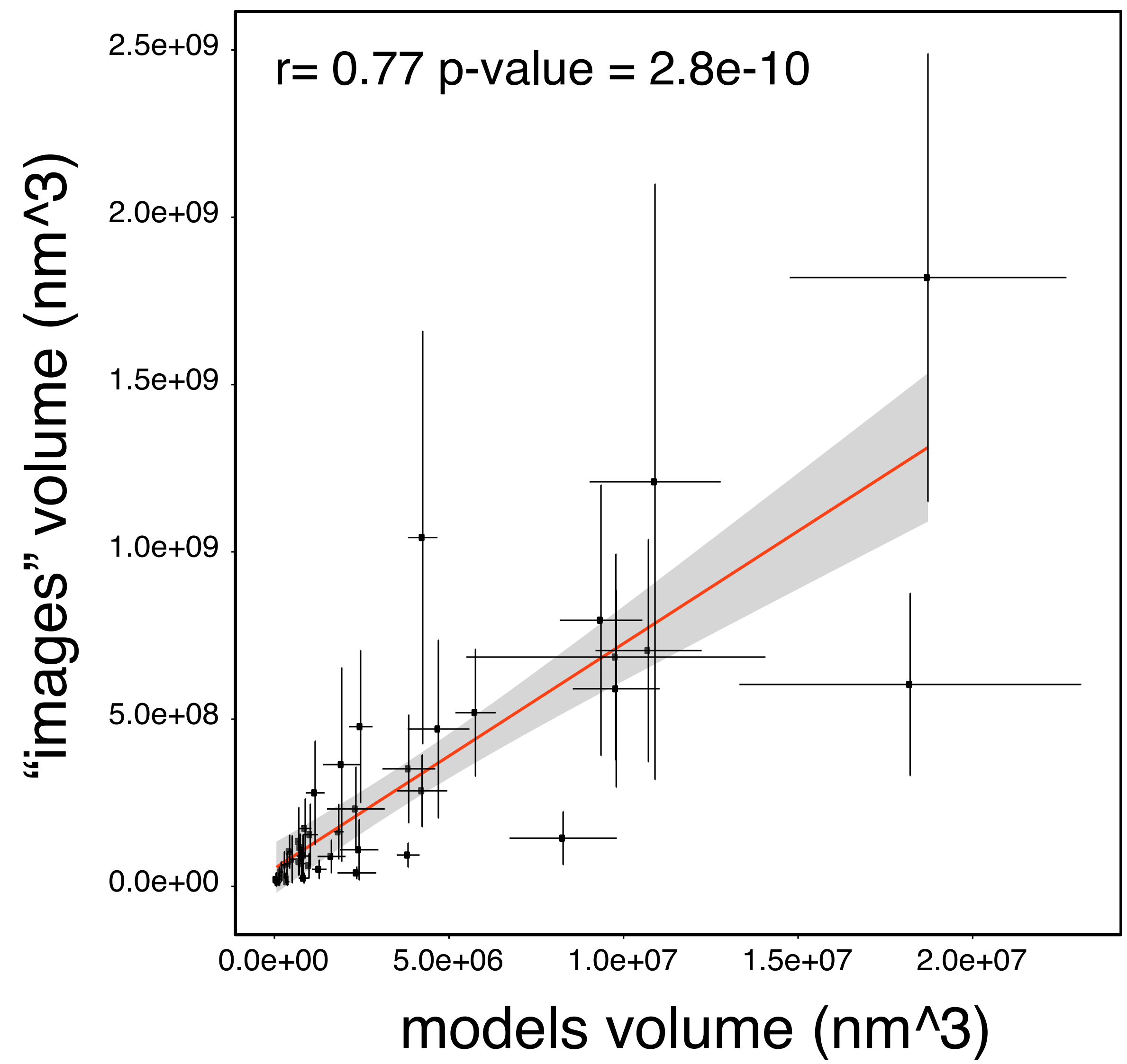
Boettiger, A. N., et al. (2016). Nature, 529, 418–422.



Model accuracy (fly@2Kb)

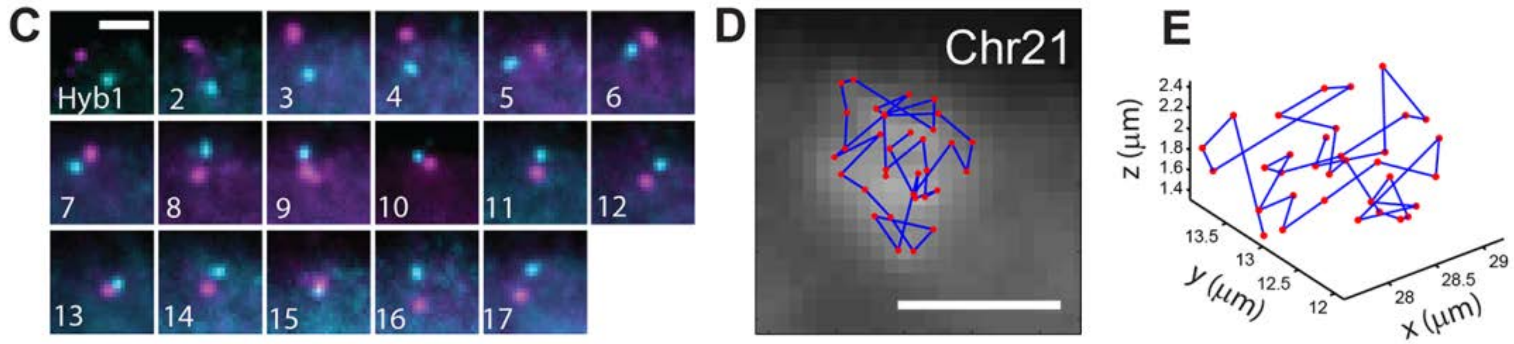
Boettiger, A. N., et al. (2016). Nature, 529, 418–422.

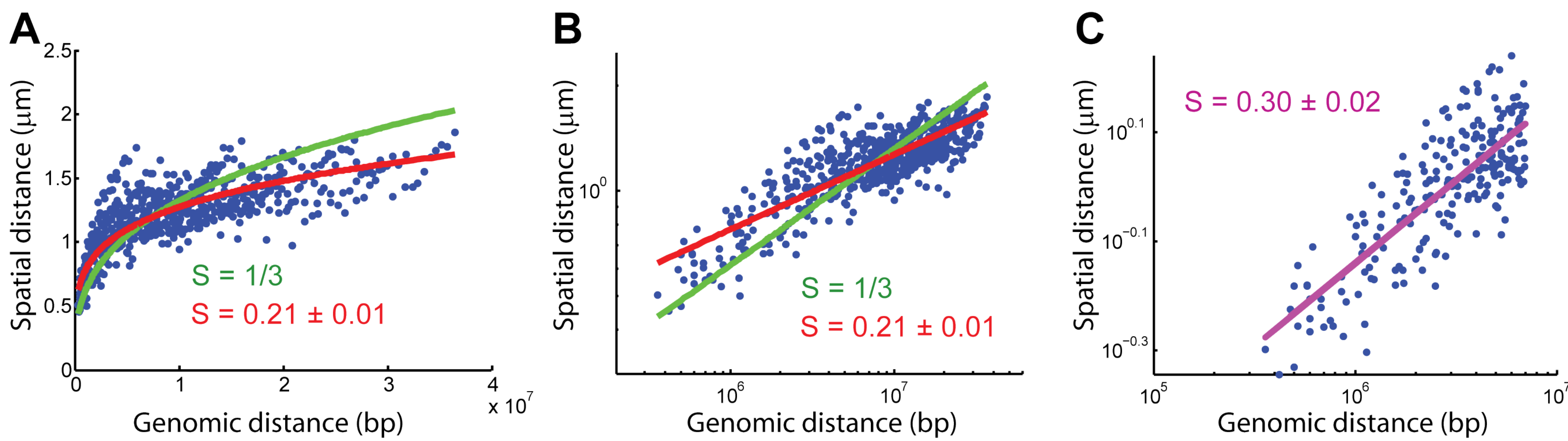


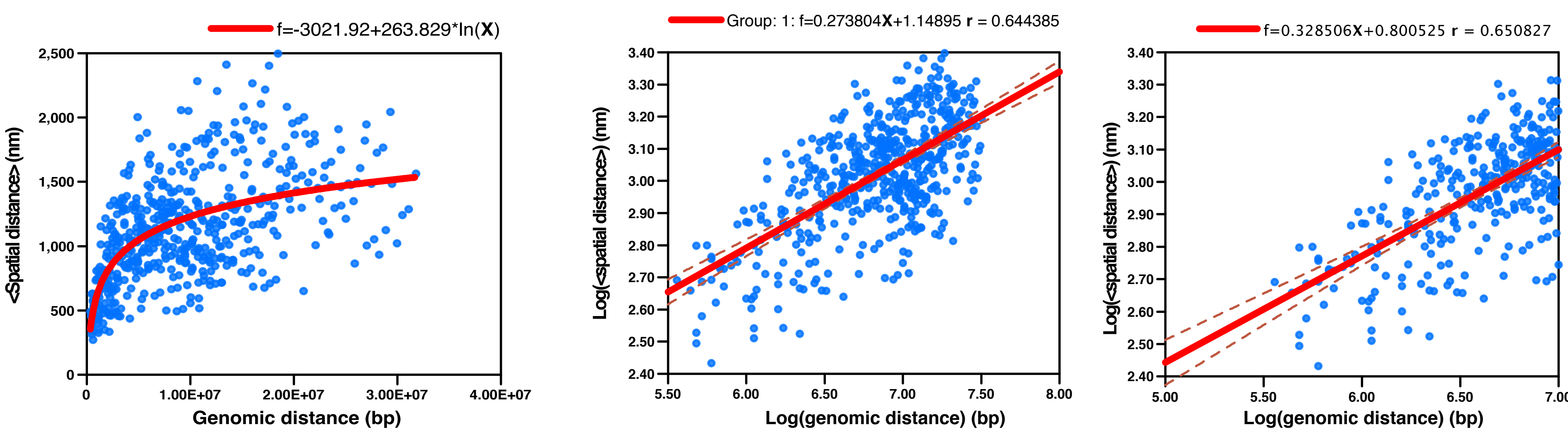


Model accuracy (Human Chr21@40Kb)

Wang, S., et al. (2016). Science 353, 598–602.







Model accuracy (Human Chr21@40Kb)

Wang, S., et al. (2016). Science 353(6299), 598–602.