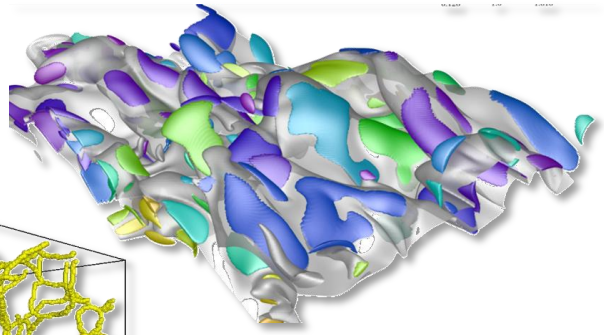
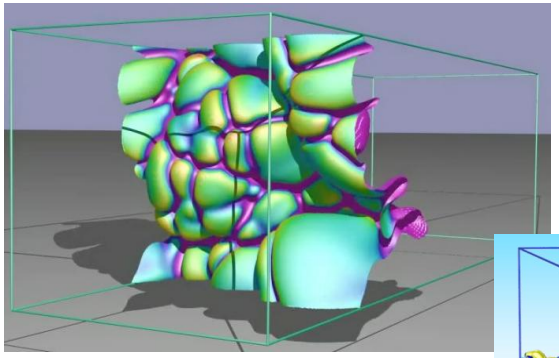
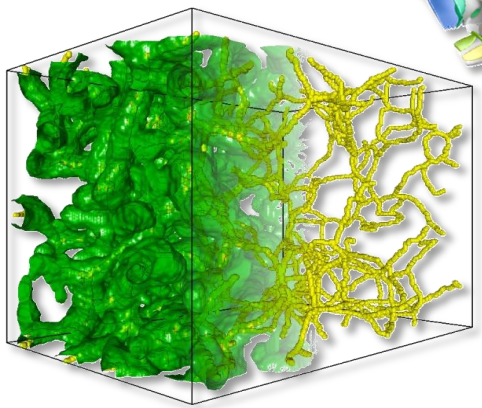
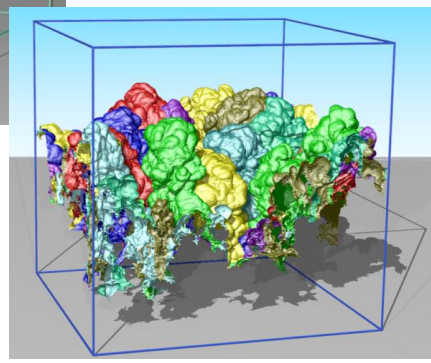
Big Data Analytics for Science Discovery

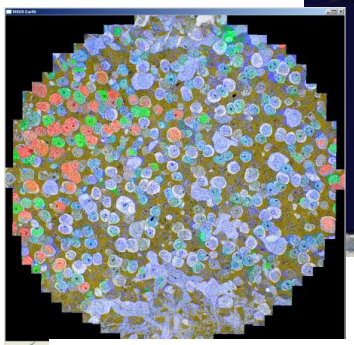
Valerio Pascucci

Director, Center for Extreme Data Management Analysis and Visualization
Professor, SCI institute and School of Computing, University of Utah
Laboratory Fellow, Pacific Northwest National Laboratory

Massive Simulation and Sensing Devices Generate Great Challenges and Opportunities


BlueGene/L


Satellite

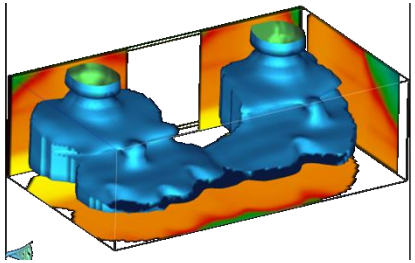

Retinal Connectome

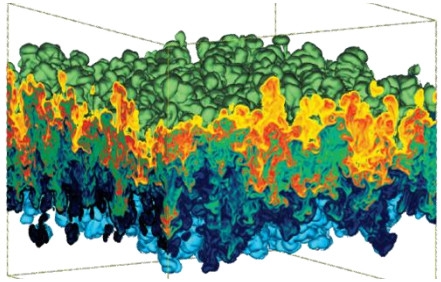

Earth Images

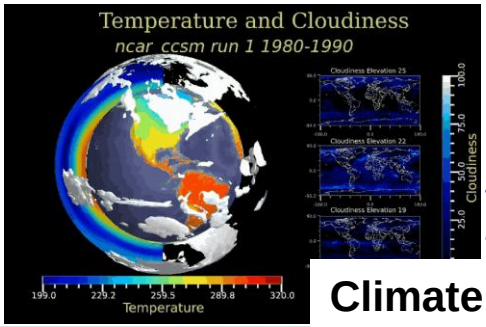

Cameras


EM

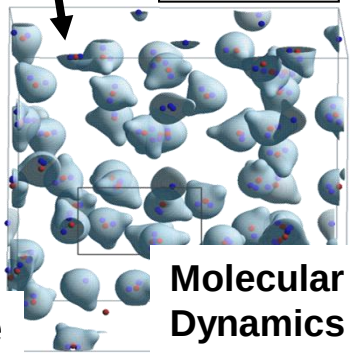
Jaguar

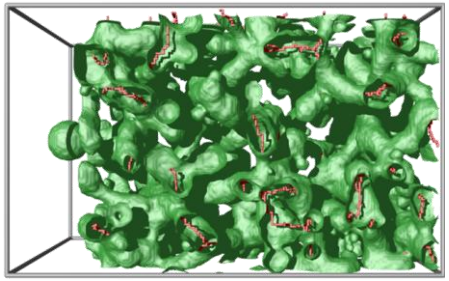

Carbon Seq. (Subsurface)

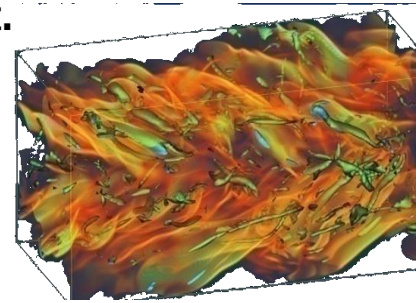

Hydrodynamic Inst.



Climate


Molecular Dynamics


Porous Materials

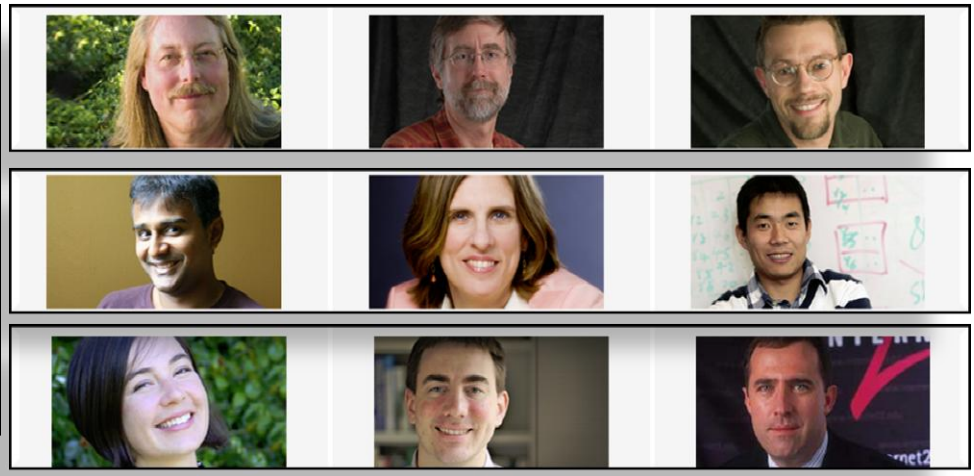

Turbulent Combustion


Photography

CEDMAV

Center for Extreme Data Management, Analysis, and Visualization

- 10 Faculty + scientists, developers, students, …
- Primary partners: UU & PNNL
- Other partnerships: NSA, INL, LLNL, ANL, Battelle, ….
- Involvement in national Initiatives



SDAV Scalable Data Management, Analysis, and Visualization

EXACT

SUPER

The University of Utah Pacific Northwest National Laboratory

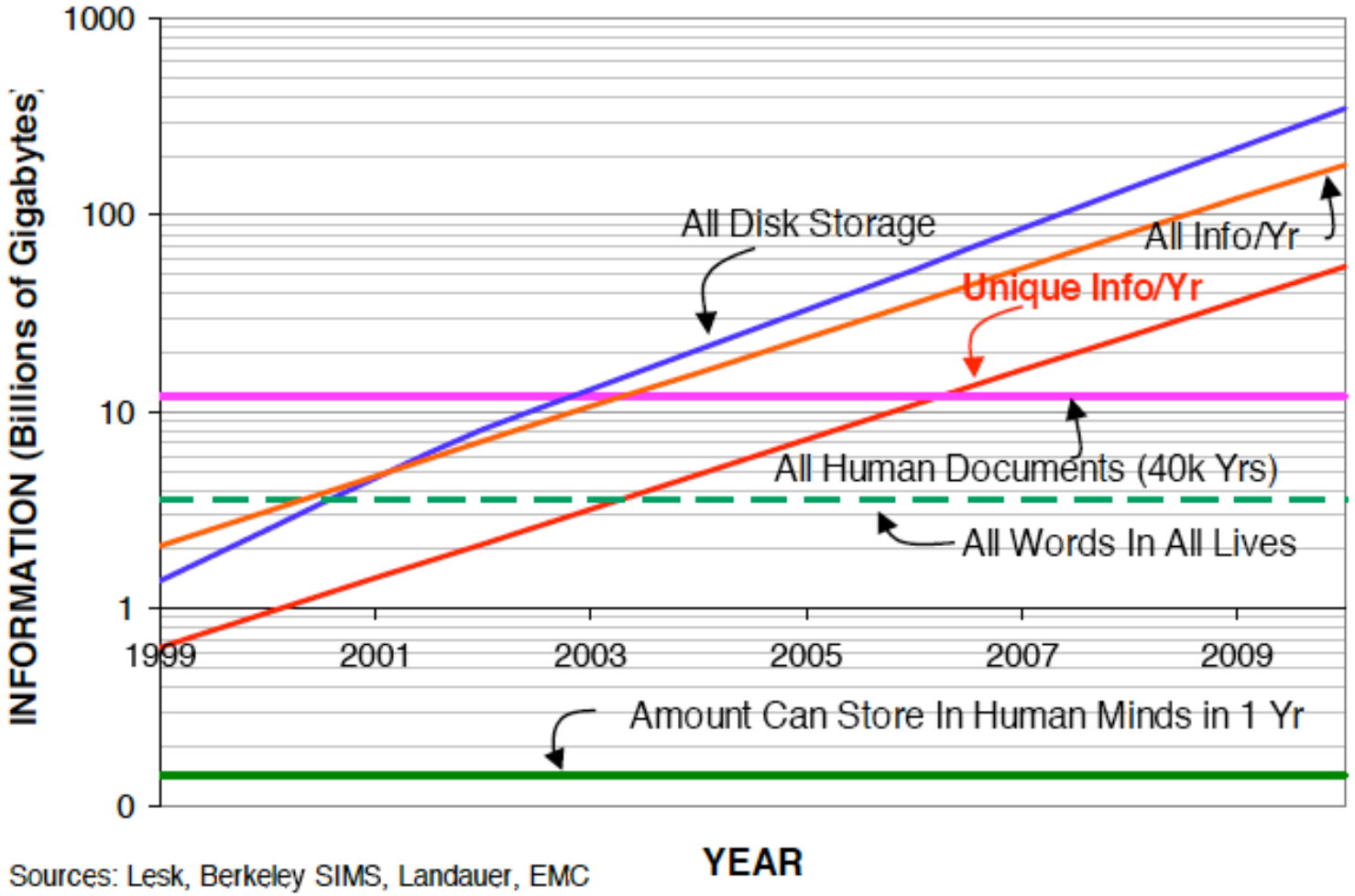
Lawrence Livermore National Laboratory

Sandia National Laboratories

INL Idaho National Laboratory

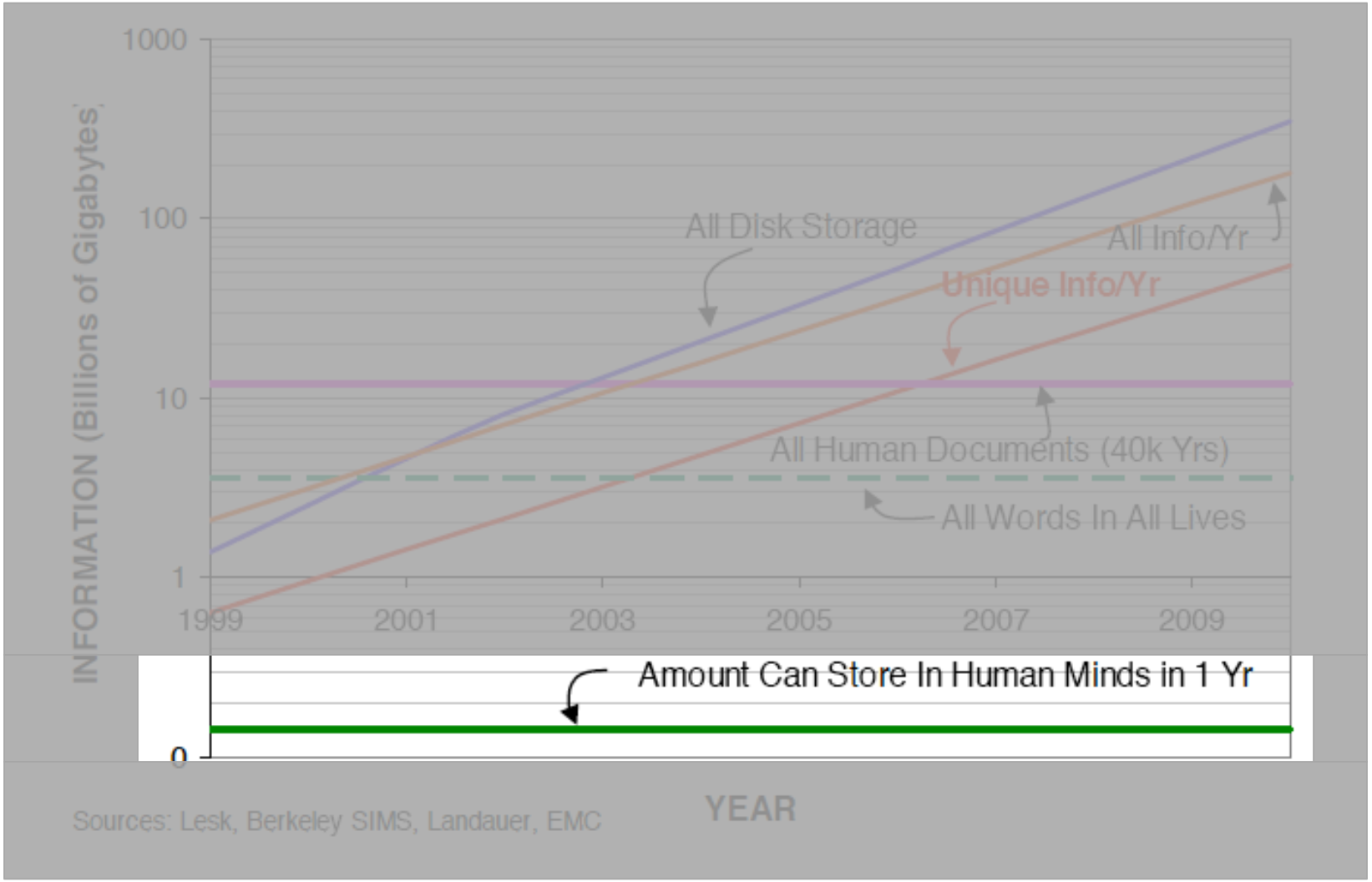
$1.6B NSA data center
(1.5 million-square-foot facility)

The Information Big Bang Has Come!



Sources: Lesk, Berkeley SIMS, Landauer, EMC

Are We Hopeless ?



How Much Information Is This?

01100101

00111010

11011010

11000110

00110010

How Much Information Is This?

E

00111010

11011010

11000110

00110010

How Much Information Is This?

E

=

11011010

11000110

00110010

How Much Information Is This?

E

=

m

11000110

00110010

How Much Information Is This?

E

=

m

c

00110010

How Much Information Is This?

$$E = mc^2$$

How Much Information Is This?

$$E=mc^2$$

Multiresolution vs Abstraction

Multiresolution vs Abstraction

Multiresolution vs Abstraction

Multiresolution vs Abstraction

Multiresolution vs Abstraction

Multiresolution vs Abstraction

Multiresolution vs Abstraction

Multiresolution vs Abstraction

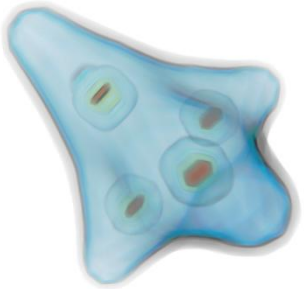
Multiresolution vs Abstraction

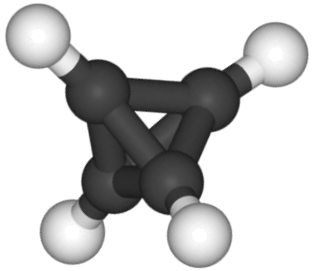
Multiresolution vs Abstraction

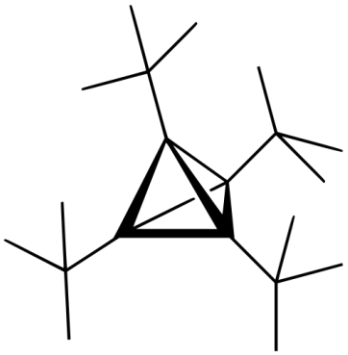
FLOWERS

Abstractions Have Been Used in Science Discovery/Communication for a Long Time

Abstraction





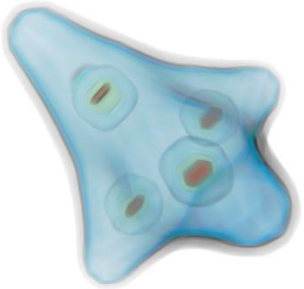


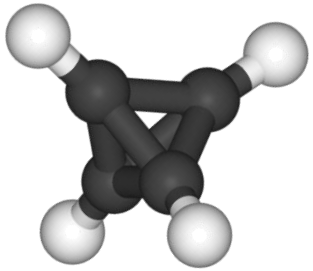
C_4H_4

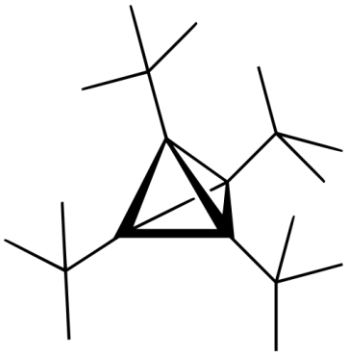
Tetrahedrane

Abstractions Have Been Used in Science Discovery/Communication for a Long Time

Abstraction



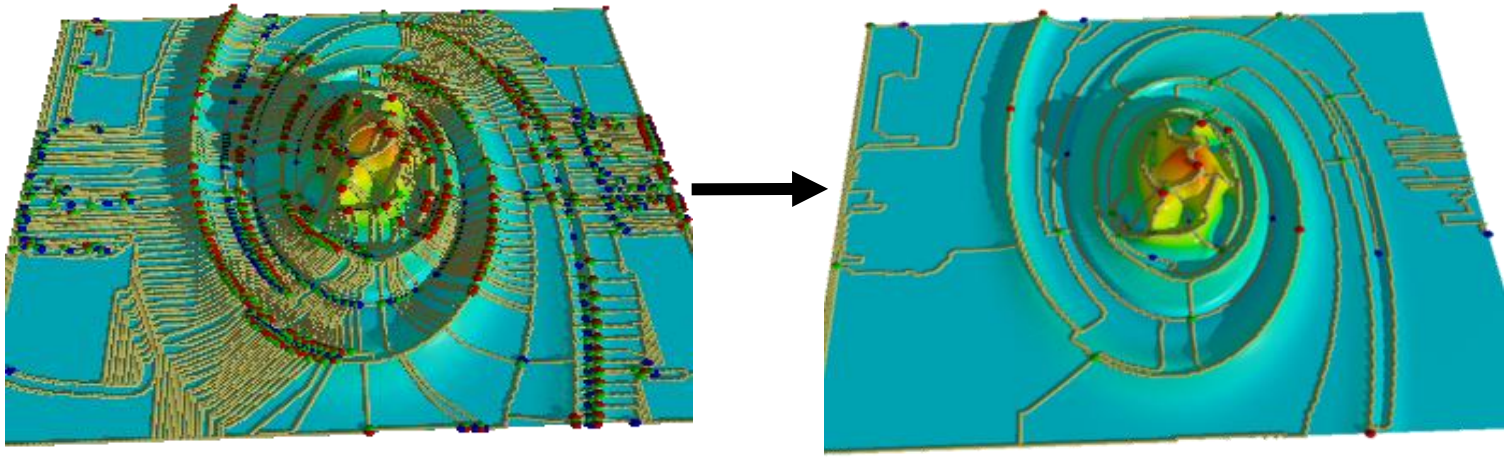




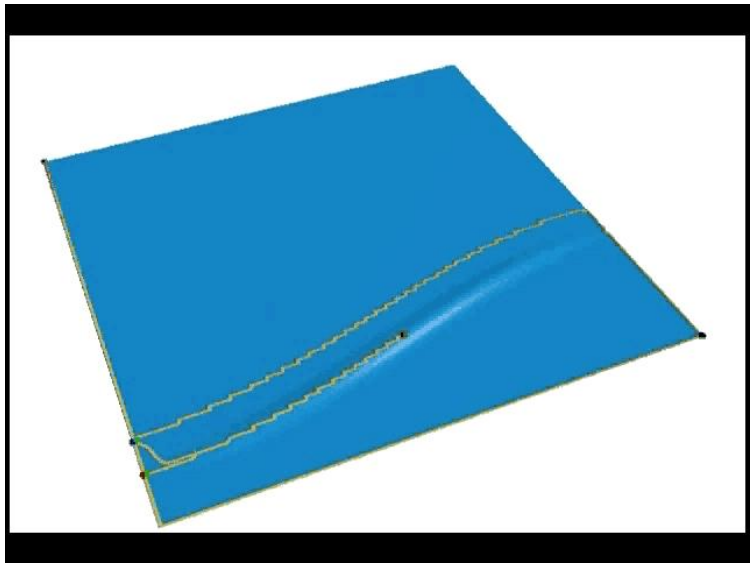
C_4H_4 Tetrahedrane

Language

Topology is and Effective Language to Describe Abstractions of Features from Raw Data

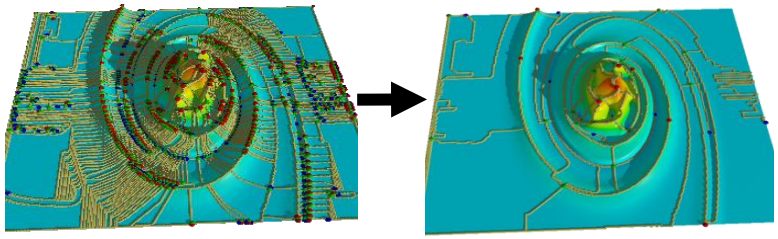


Hierarchical topology of a 2D Miranda vorticity field

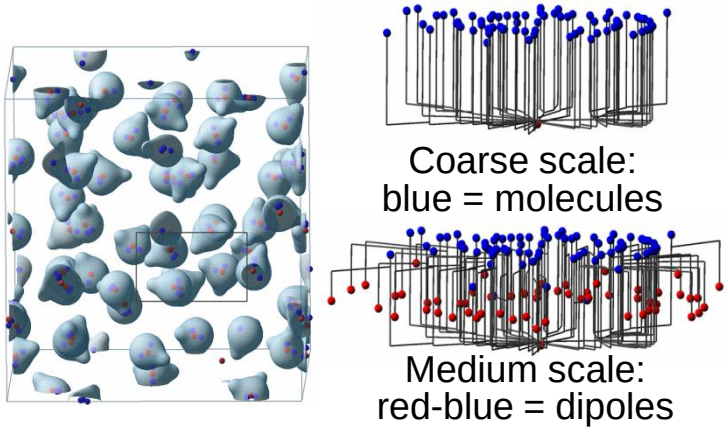


Our Framework is Based on Robust Topological Computations for Quantitative Data Analysis

- **Provably robust computation**
- **Provably complete feature extraction and quantification**
- **Hierarchical topological structures used to capture multiple scales**
- **Error-bounded approximations associated with each scale**
- **Formal mathematical definition associated with each analysis**
- **Scalable performance in association with streaming techniques**



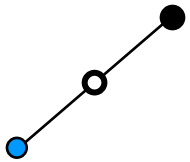
Hierarchical topology of a 2D Miranda vorticity field



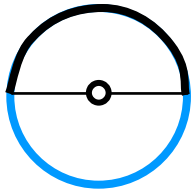
Molecular dynamics simulation (left) with abstract graph representation of its features at two scales (right)

We Adopted a Combinatorial Approach to Morse Theory for Provably Correct Computations

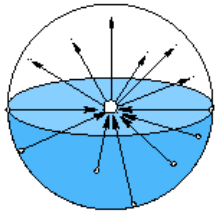
	Classical mathematical definitions	Simulation of differentiability
domain	D smooth manifold	S simplicial complex
function	f infinitely differentiable	$F(x)$ PL-extension of $f(x_i)$
critical point	**numerical**	**combinatorial**



1D



2D

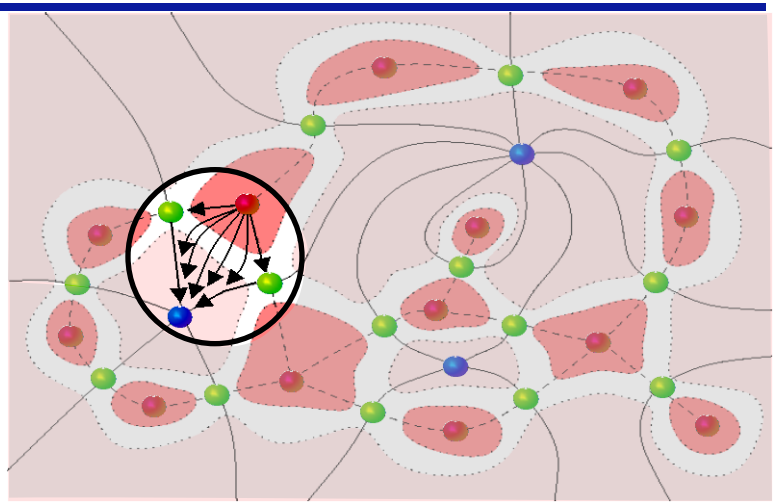


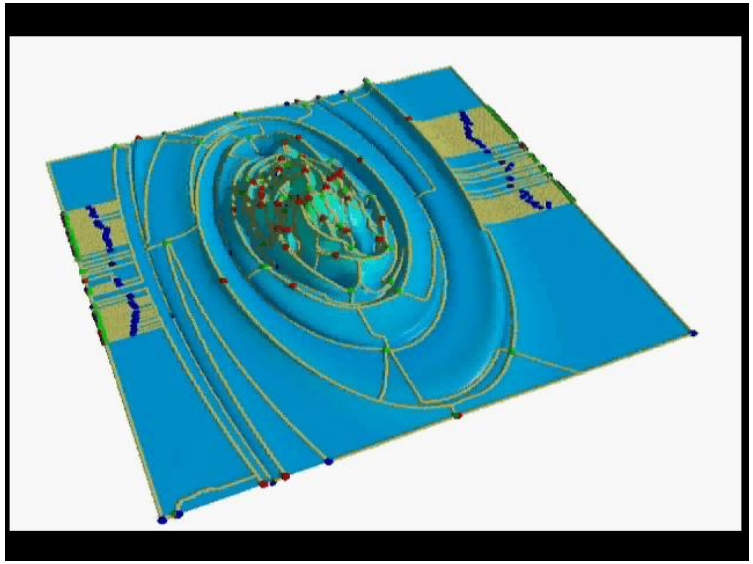
3D

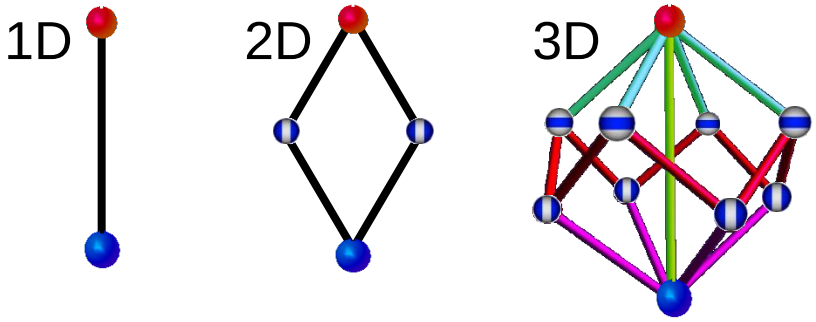
Independent local computation yield globally consistent results

We Introduced the Morse–Smale Complex for Complete Data Analysis

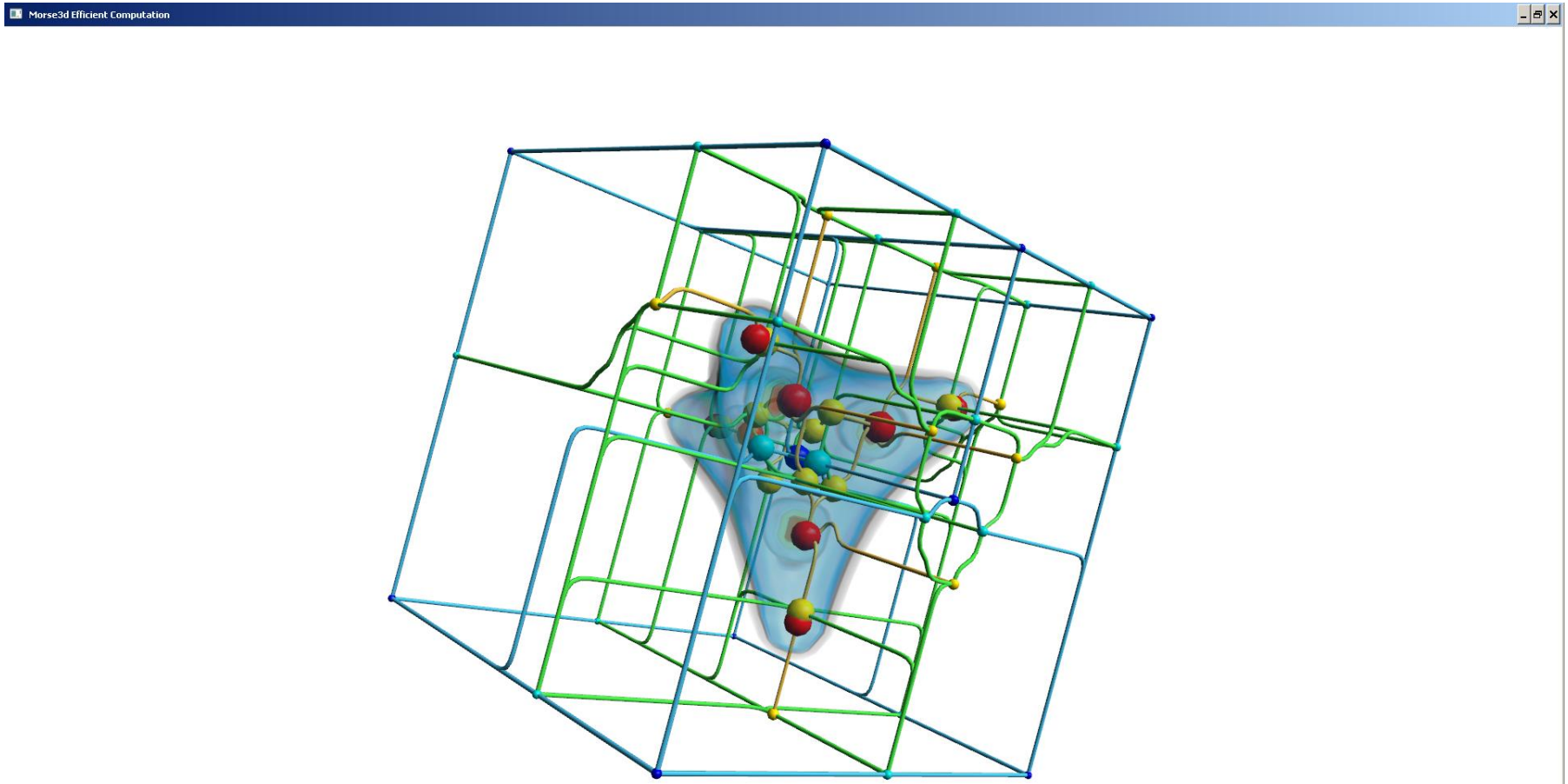
- **The Morse–Smale complex partitions the domain of *f* in regions of uniform gradient**
- **Generalizes the notion of monotonic interval**
- **Dimension of a region equal index difference of source and destination**
- **Remove inconsistency of local gradient evaluations**







Demo C_4H_4

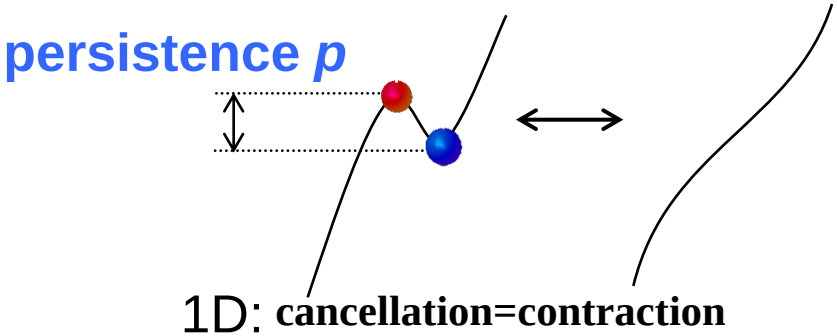


We Use Cancellations to Create a Multi-scale Representation of the Trends in the Data

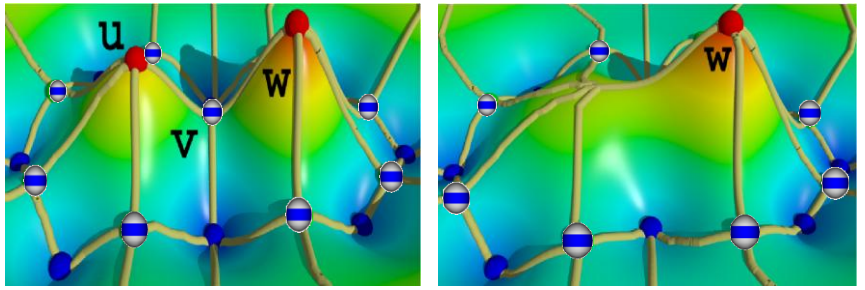
Cancellations: critical points can be created or destroyed in pairs that are connected 1-manifolds

Approximation: error = persistence/2 (proven lower bound)

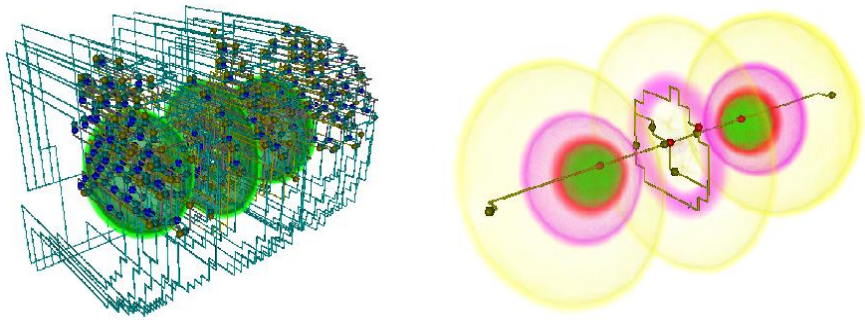
Multi-scale: consistent gradient segmentation at all scales



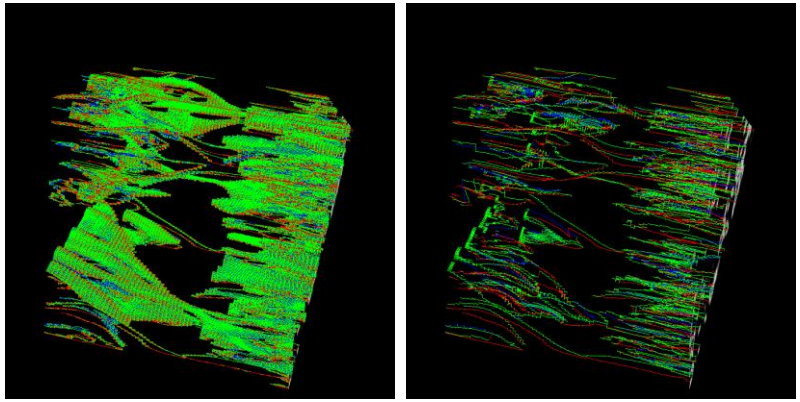
1D: **cancellation=contraction**



2D: **cancellation=contraction + edge removal**

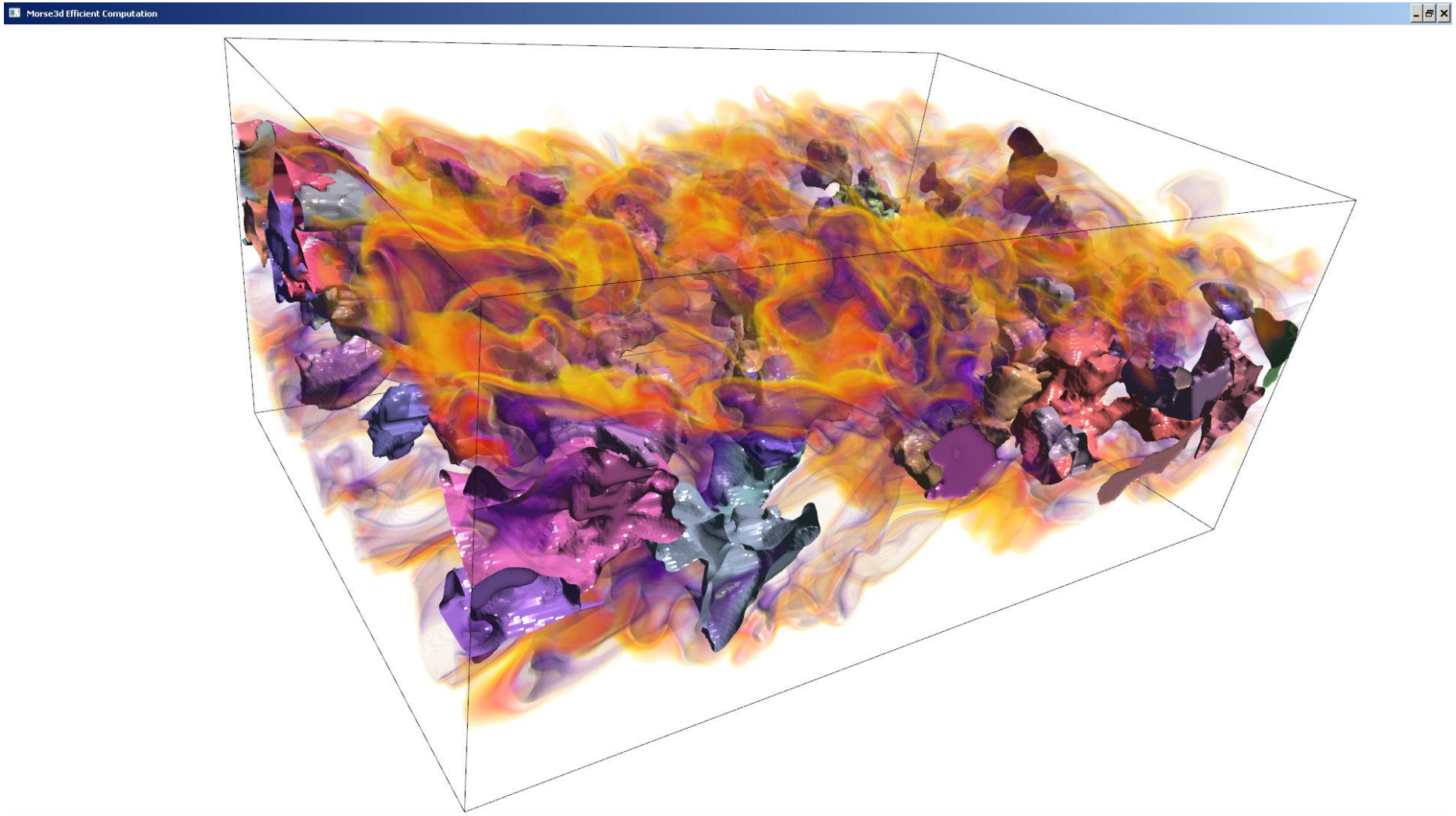


3D

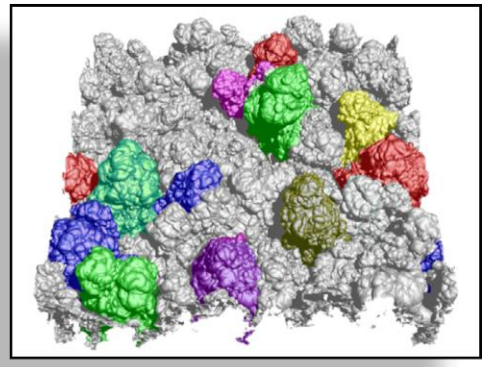


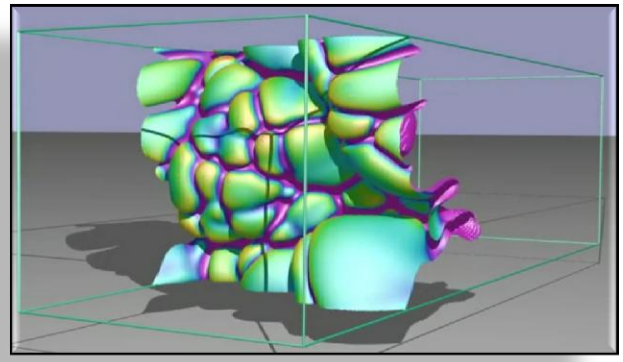
Time

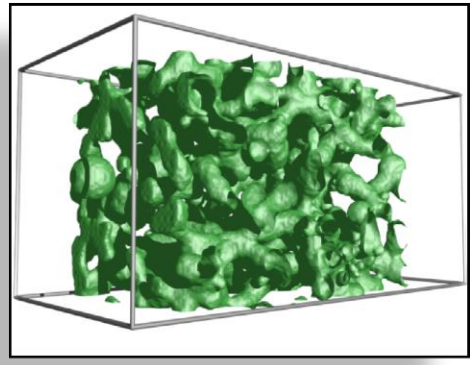
Demo S3D Combustion Simulation

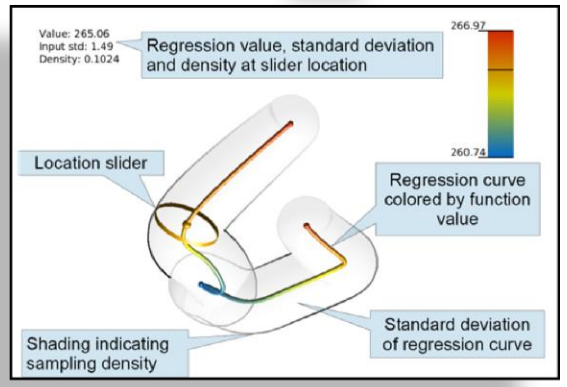


Big Data Analytics Success Stories

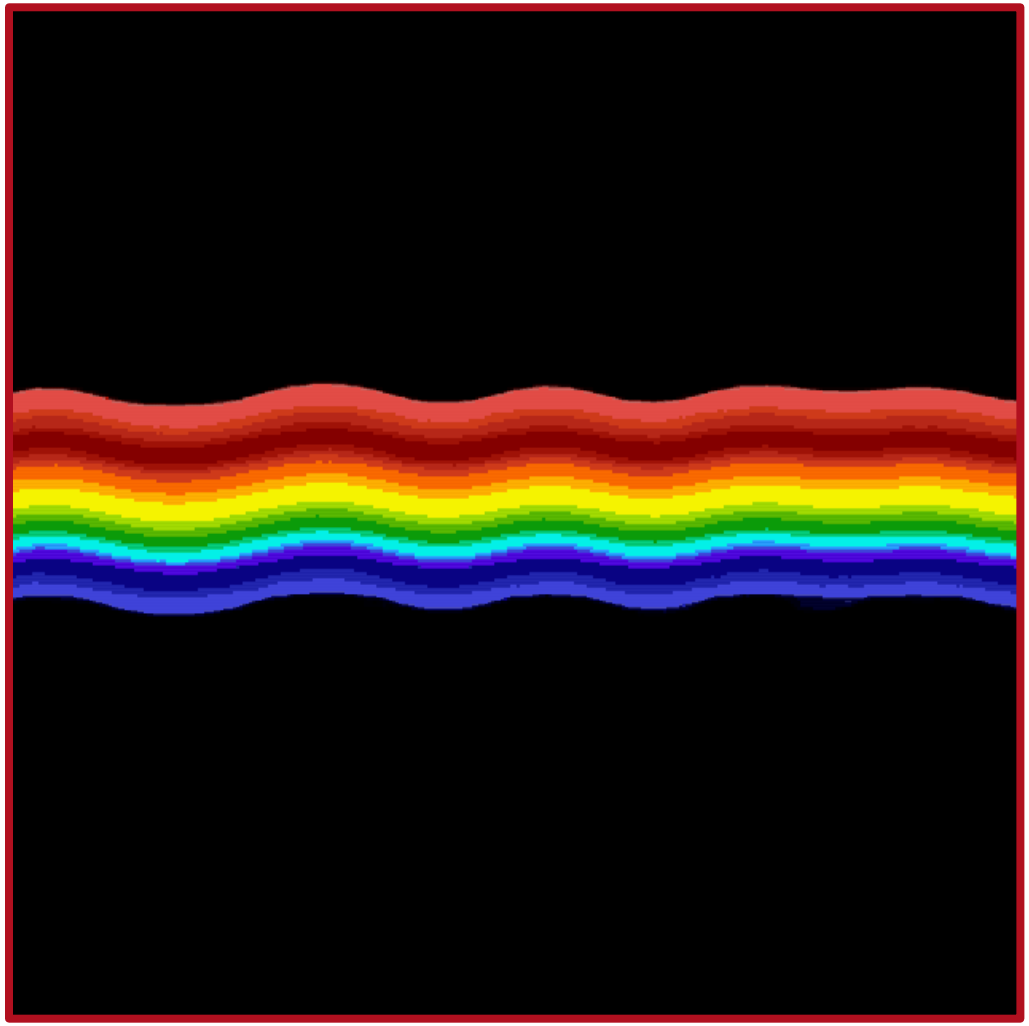








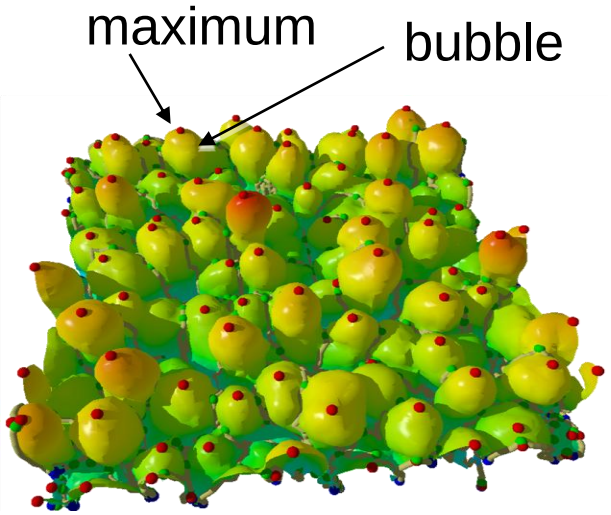
Count the Number of Bubbles in a Rayleigh–Taylor Instability



Rayleigh-Taylor instabilities arise in fusion, super-novae, and other fundamental phenomena:

- start: heavy fluid above, light fluid below
- gravity drives the mixing process
- the mixing region lies between the upper envelope surface (red) and the lower envelope surface (blue)
- 25 to 40 TB of data from simulations

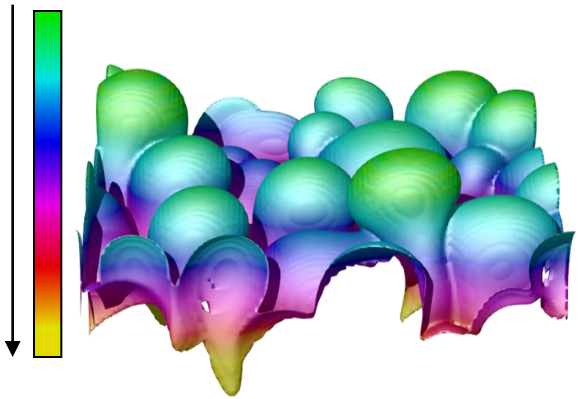
We Analyze High-Resolution Rayleigh–Taylor Instability Simulations



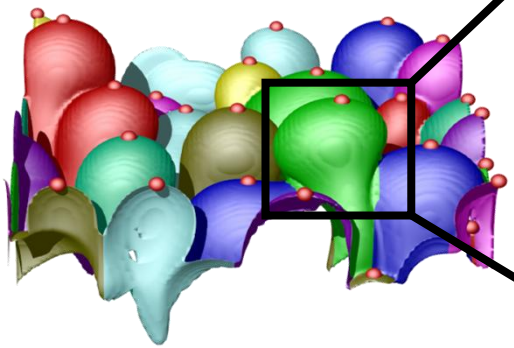
- **Large eddy simulation run on Linux cluster: 1152 x 1152 x 1152**
  - **~ 40 G / dump**
  - **759 dumps, about 25 TB**
- **Direct numerical simulation run on BlueGene/L: 3072 x 3072 x Z**
  - **Z depends on width of mixing layer**
  - **More than 40 TB**
- **Bubble-like structures are observed in laboratory and simulations**
- **Bubble dynamics are considered an important way to characterize the mixing process**
  - **Mixing rate =** $\partial(\#bubbles)/\partial t.$
- **There is no prevalent formal definition of bubbles**

We Compute the Morse–Smale Complex of the Upper Envelope Surface

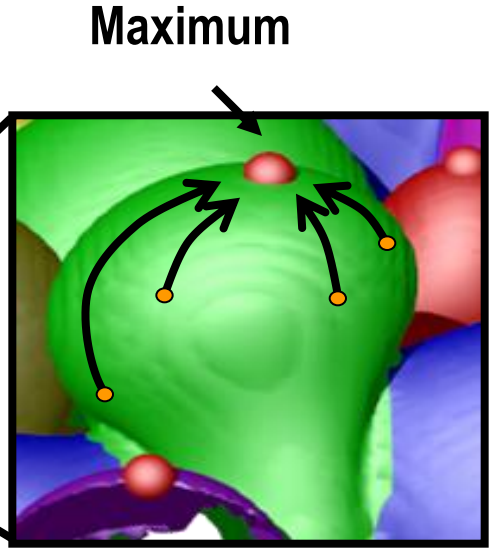
$F(x) = z$



F(x) on the surface is aligned against the direction of gravity which drives the flow



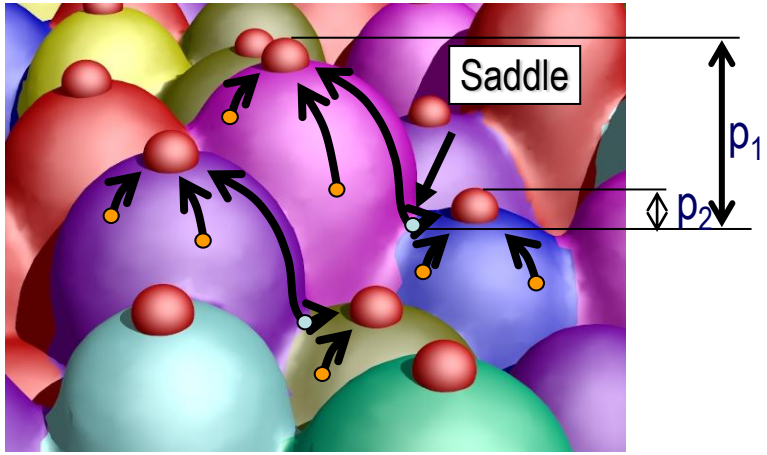
Morse complex cells drawn in distinct colors

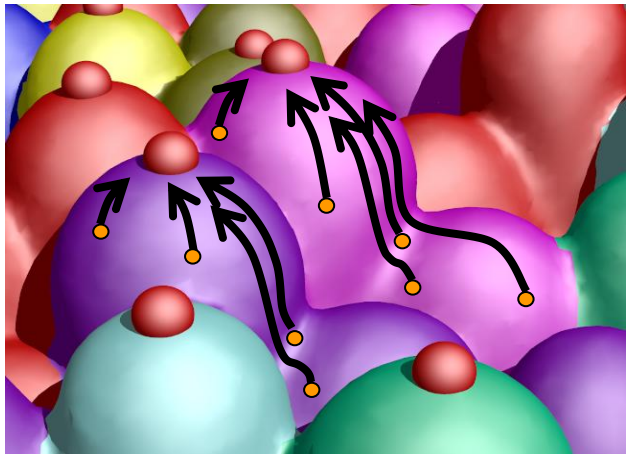


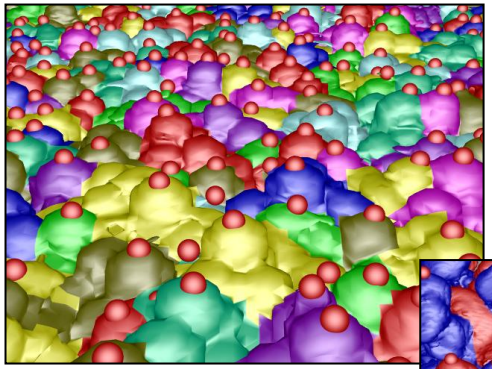
In each Morse complex cell, all steepest ascending lines converge to one maximum

A Hierarchal Model is Generated by Simplification of Critical Points

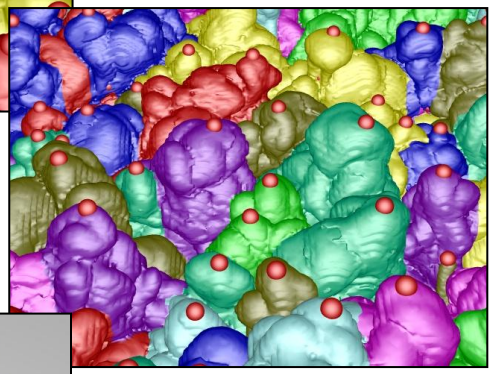
- Persistence is used to produce coarse segmentations
- Coarse scales preserve high-persistence critical points



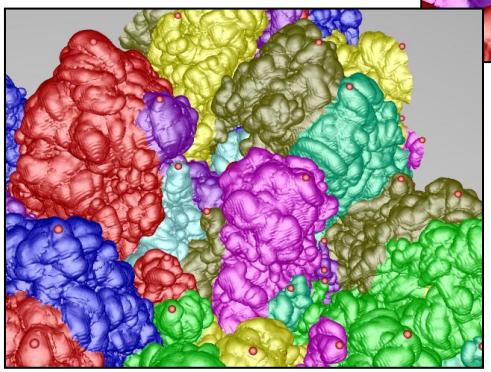




T=100

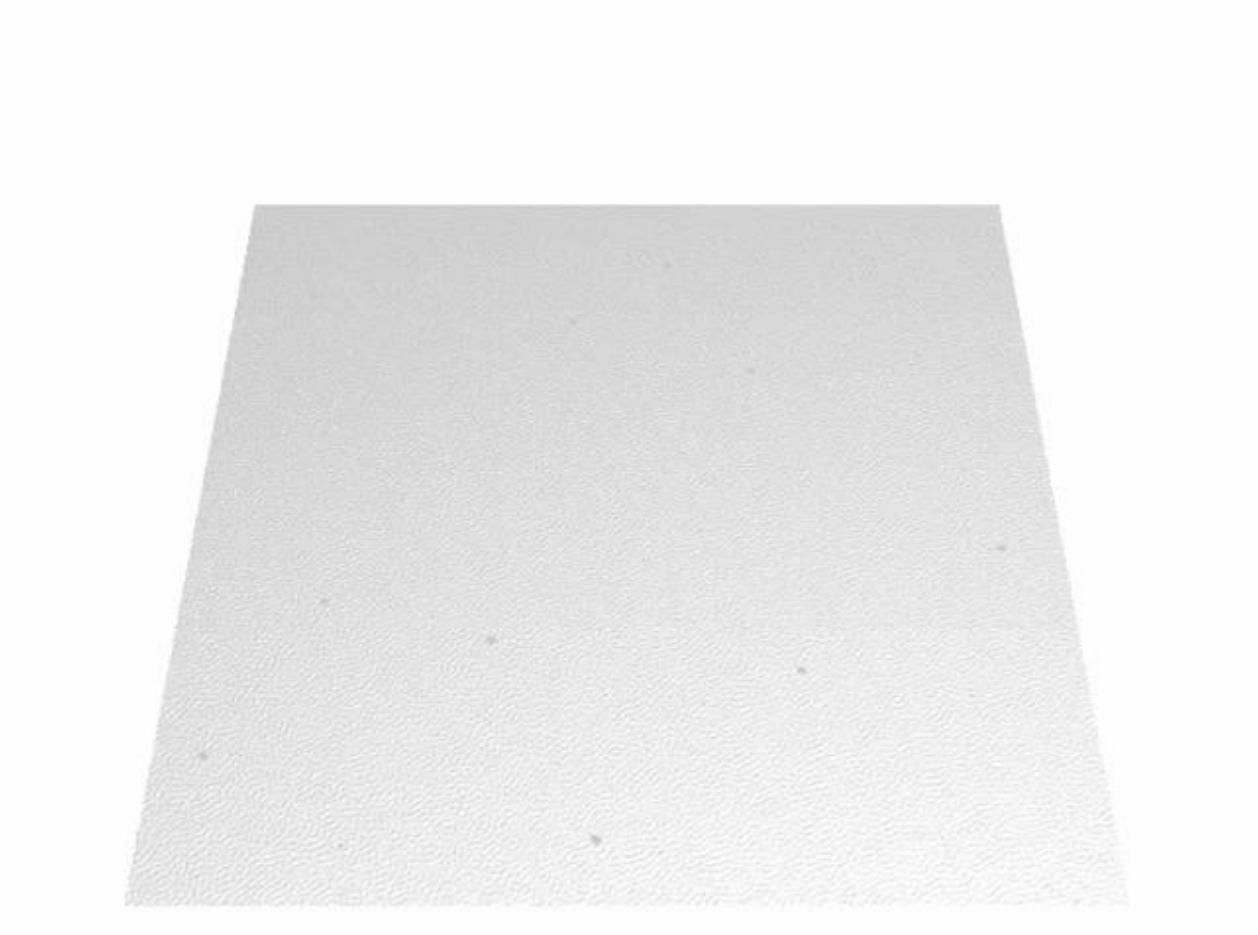


T=353

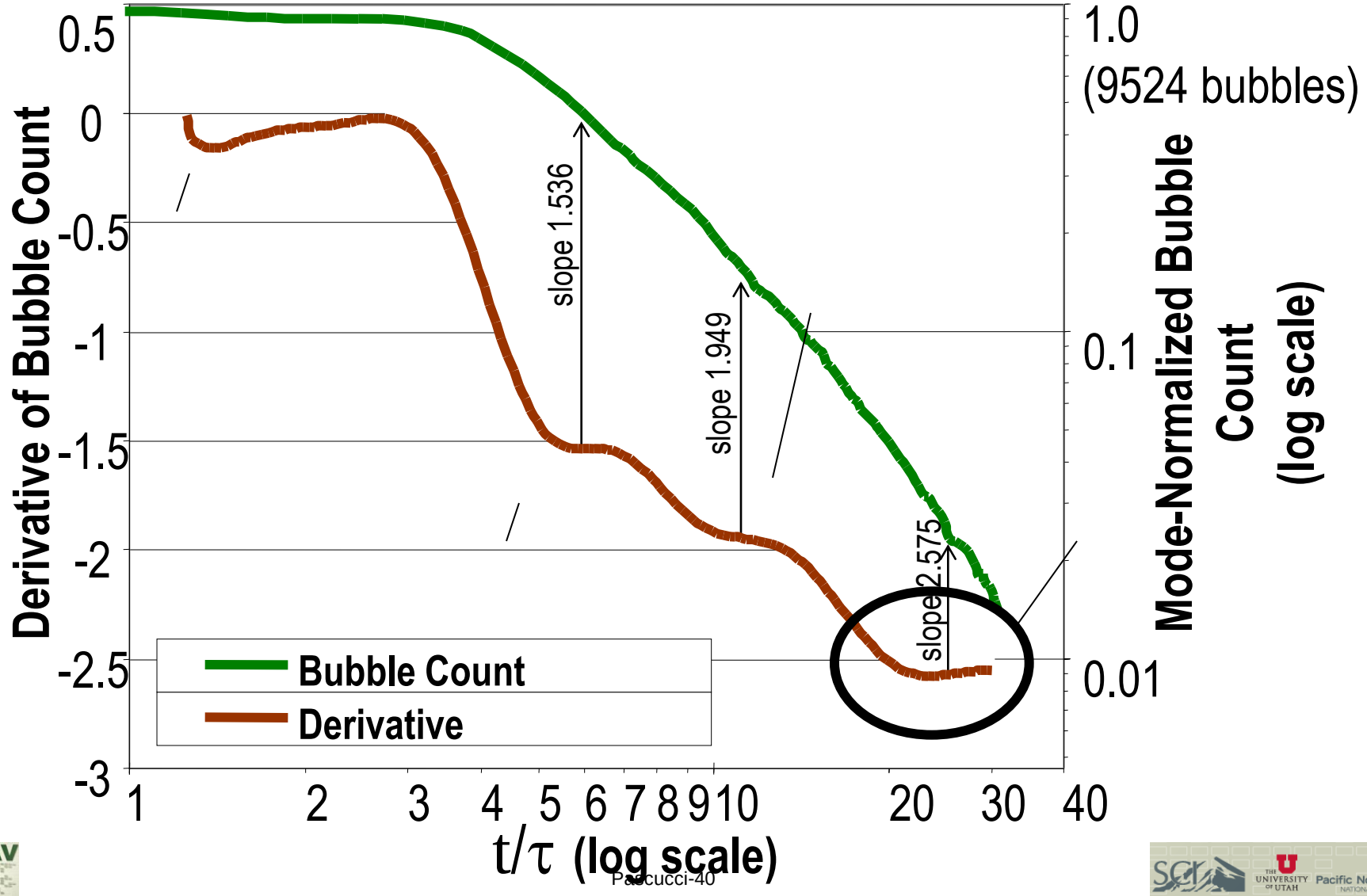


T=700

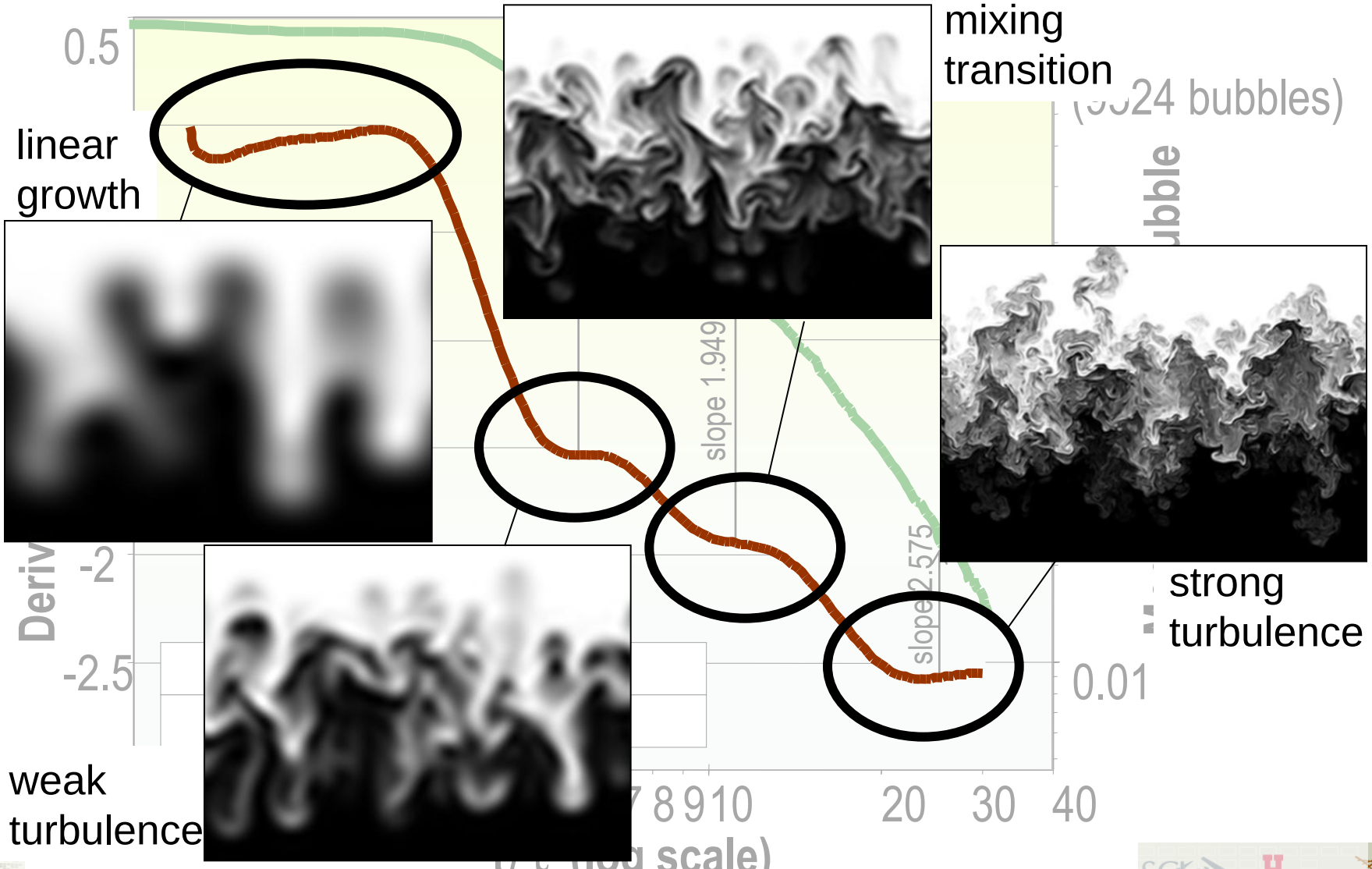
First Robust Bubble Tracking From Beginning to Late Turbulent Stages



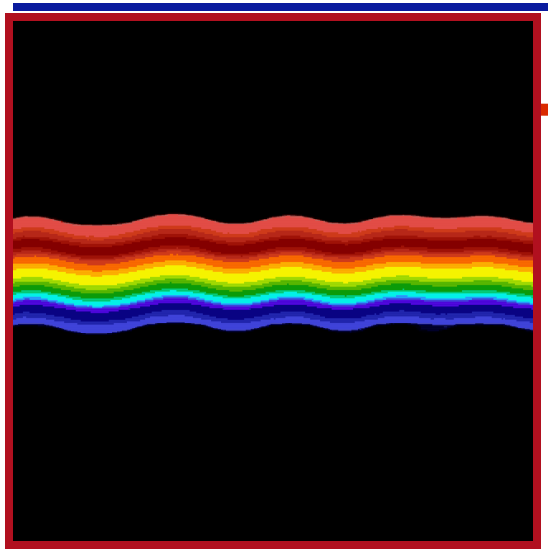
First Time Scientists Can Quantify Robustly Mixing Rates by Bubble Count



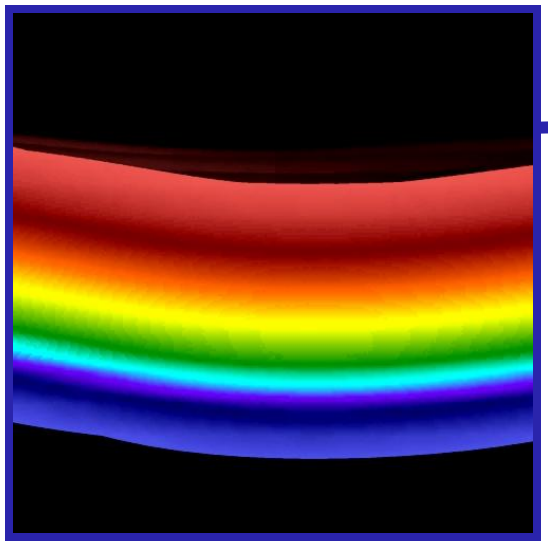
We Provide the First Quantification of Known Stages of the Mixing Process



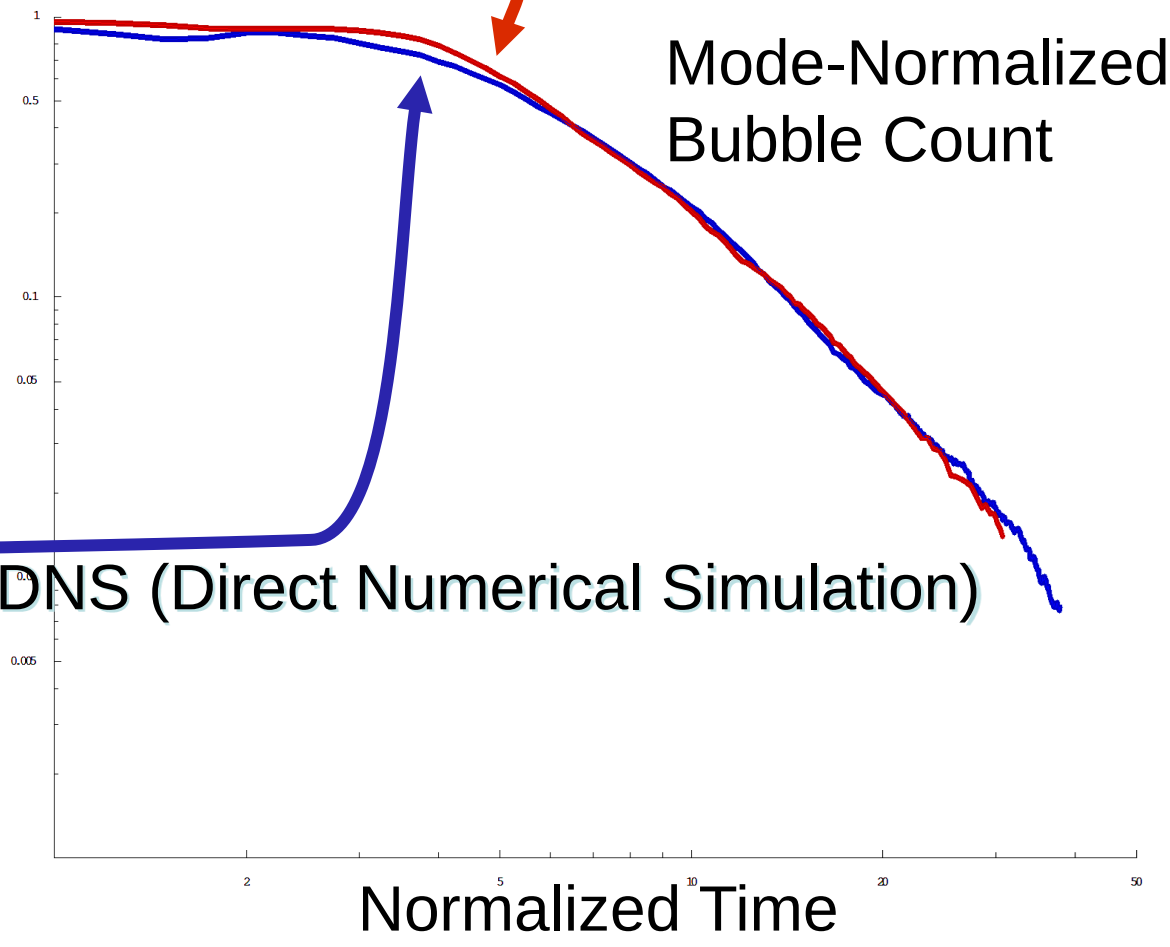
We Provided the First Feature-Based Validation of a LES with Respect to a DNS



LES (Large Eddy Simulation)

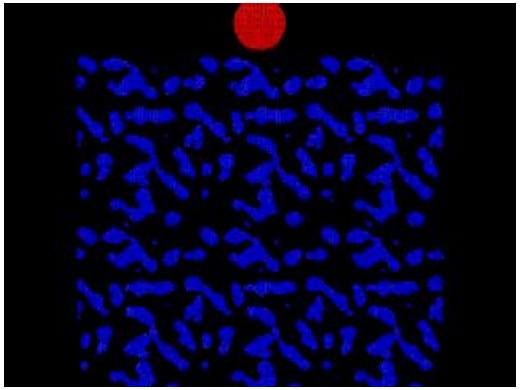


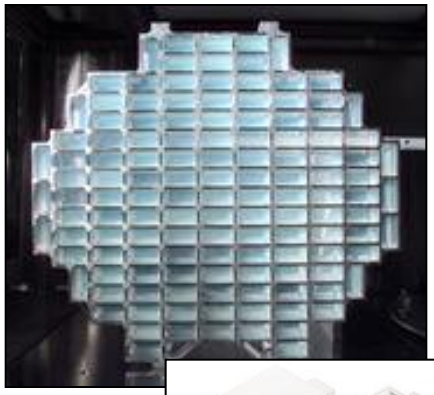
DNS (Direct Numerical Simulation)

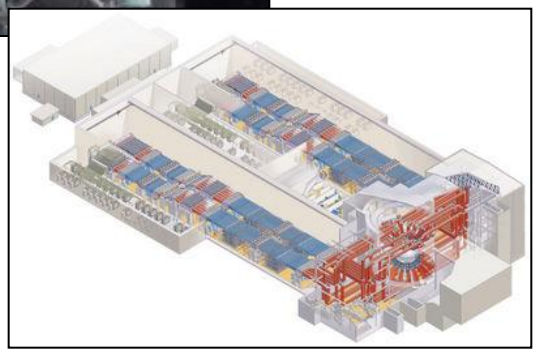


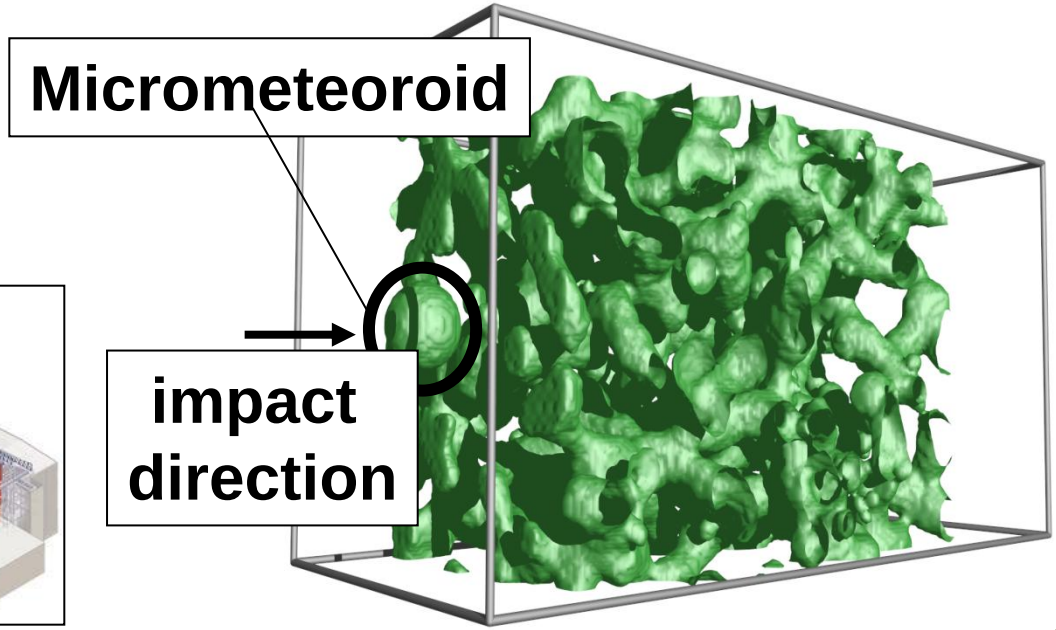
Quantitative Analysis of the Impact of a Micrometeoroid in a Porous Medium

- **Many possible applications:**
  - **NASA's Stardust Spacecraft**
  - **National Ignition Facility Targets**
  - **Light and Robust Materials**
  - **many more...**



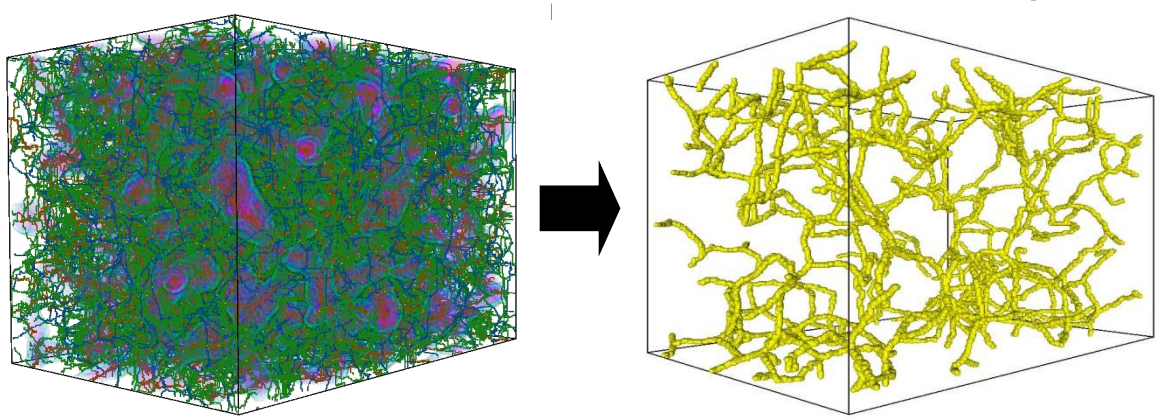




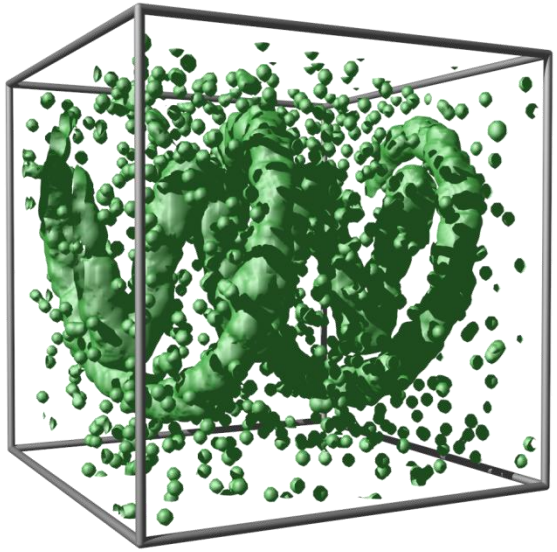


The Topological Reconstruction Method is Validated with a Controlled Test Shape

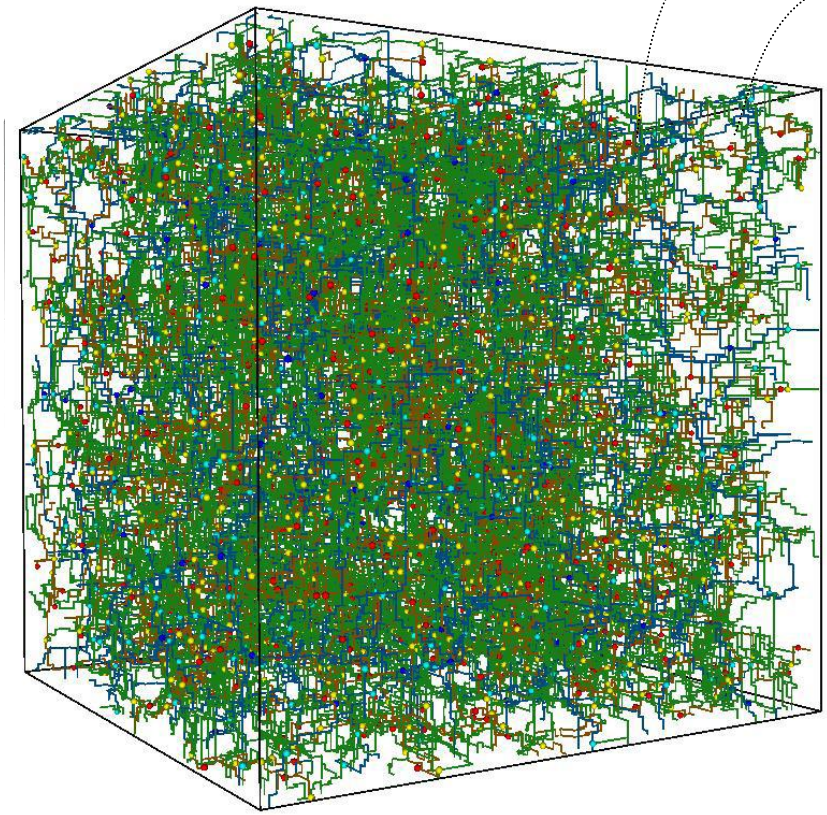
Challenge: robust reconstruction of the structure of a porous medium

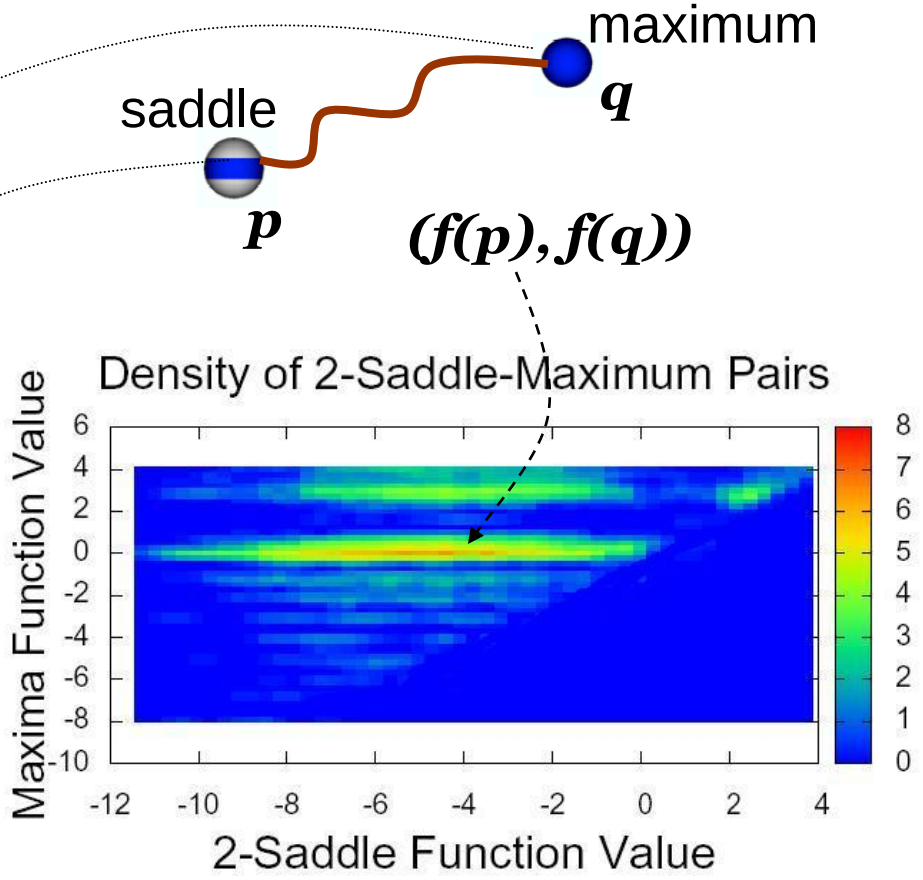


Preparation: we develop control test data to validate the approach

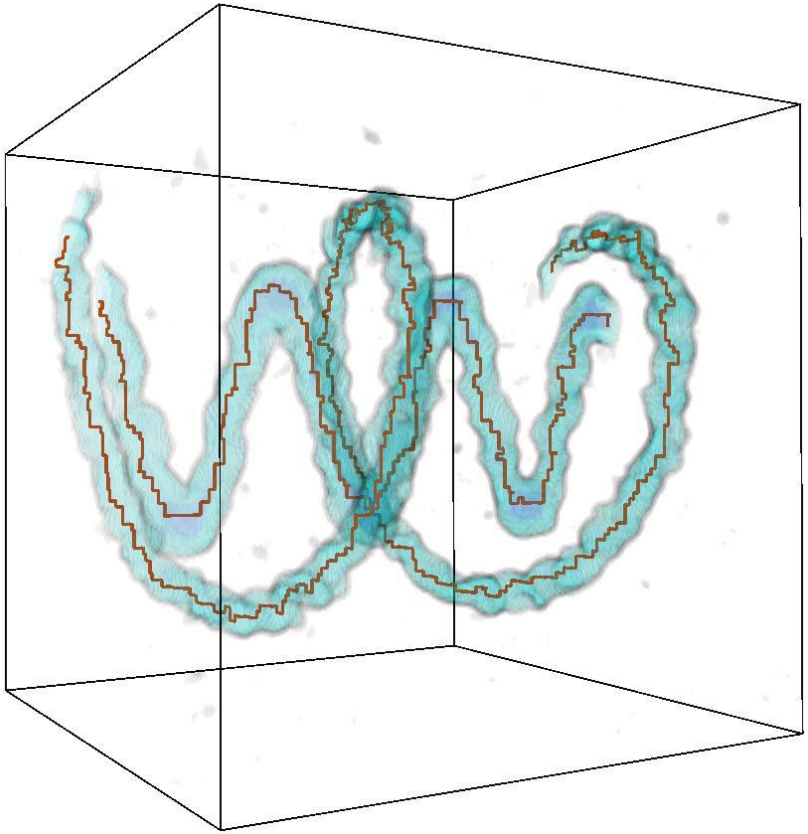


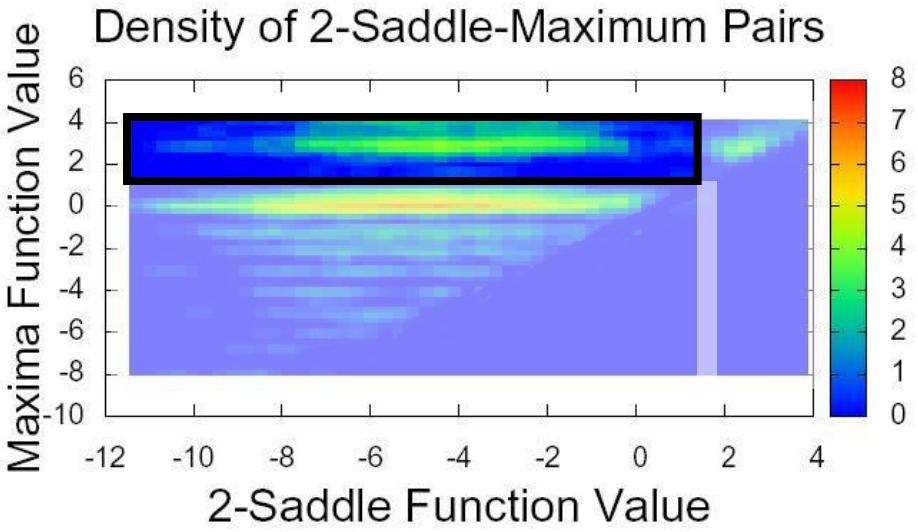
We Report the Distribution of Topological Features in the Full Resolution Data



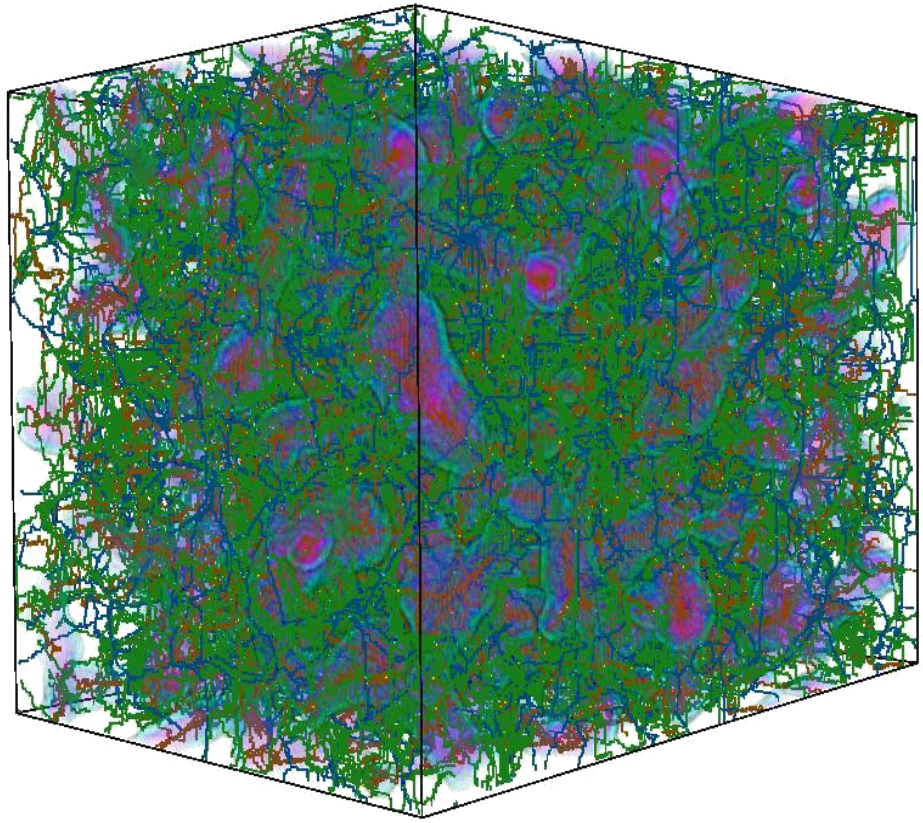


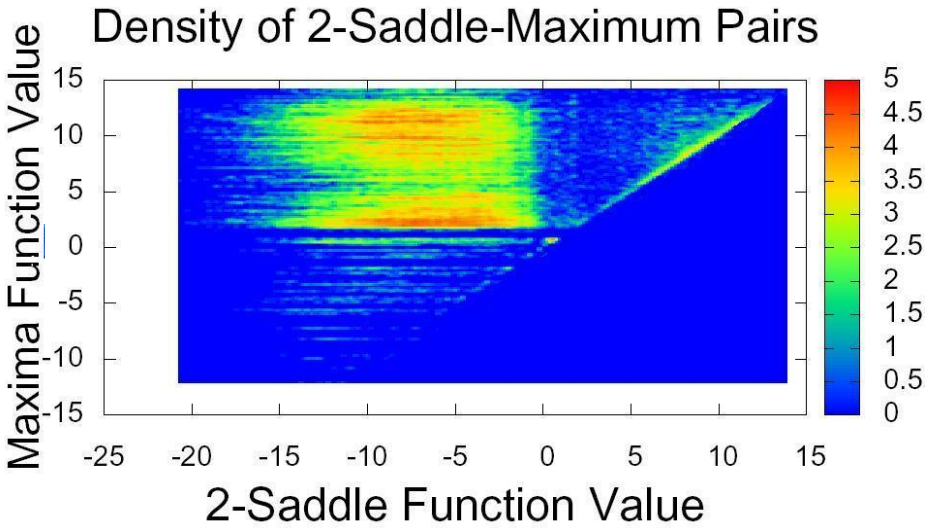
The Hierarchical Morse-Smale Complex Has Very Good Reconstruction Properties



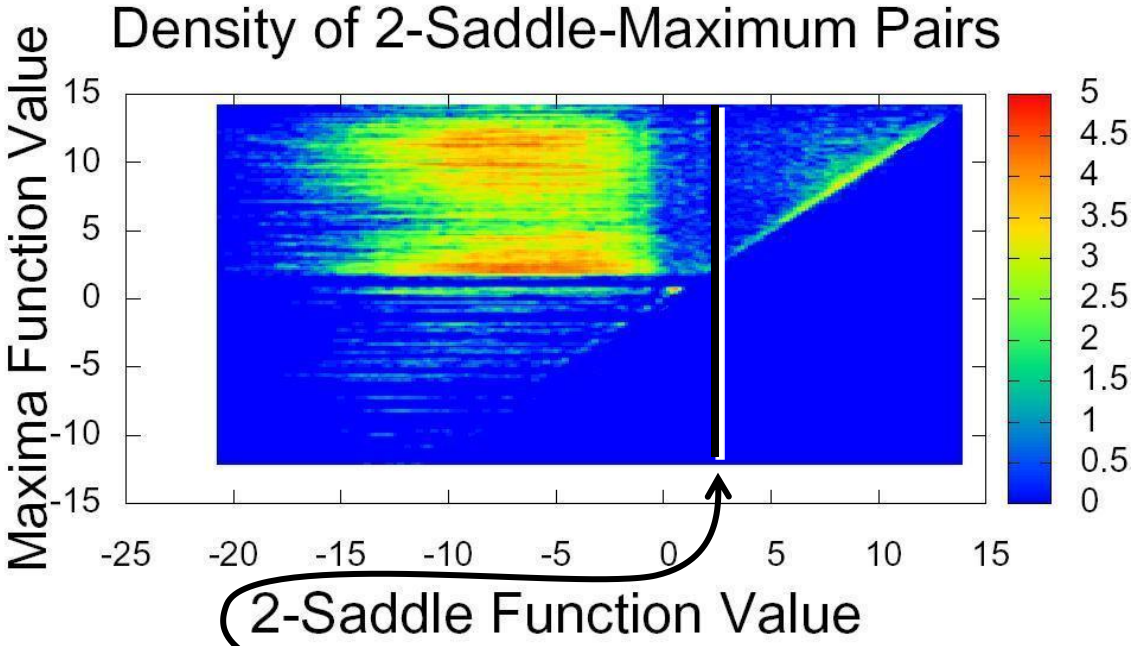


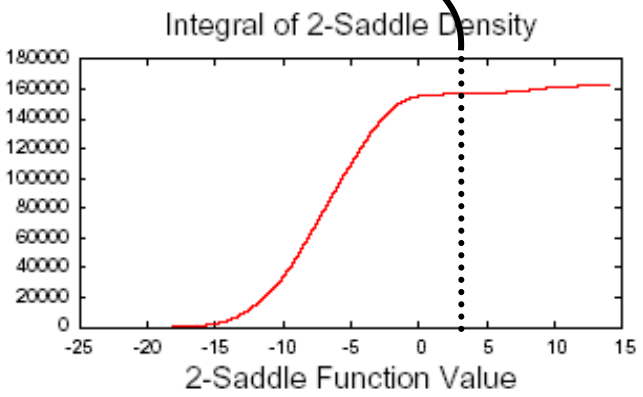
We Compute the Complete Morse-Smale Complex for the Porous Medium

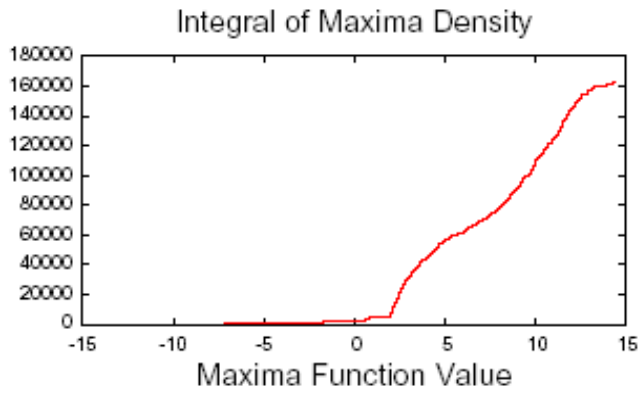




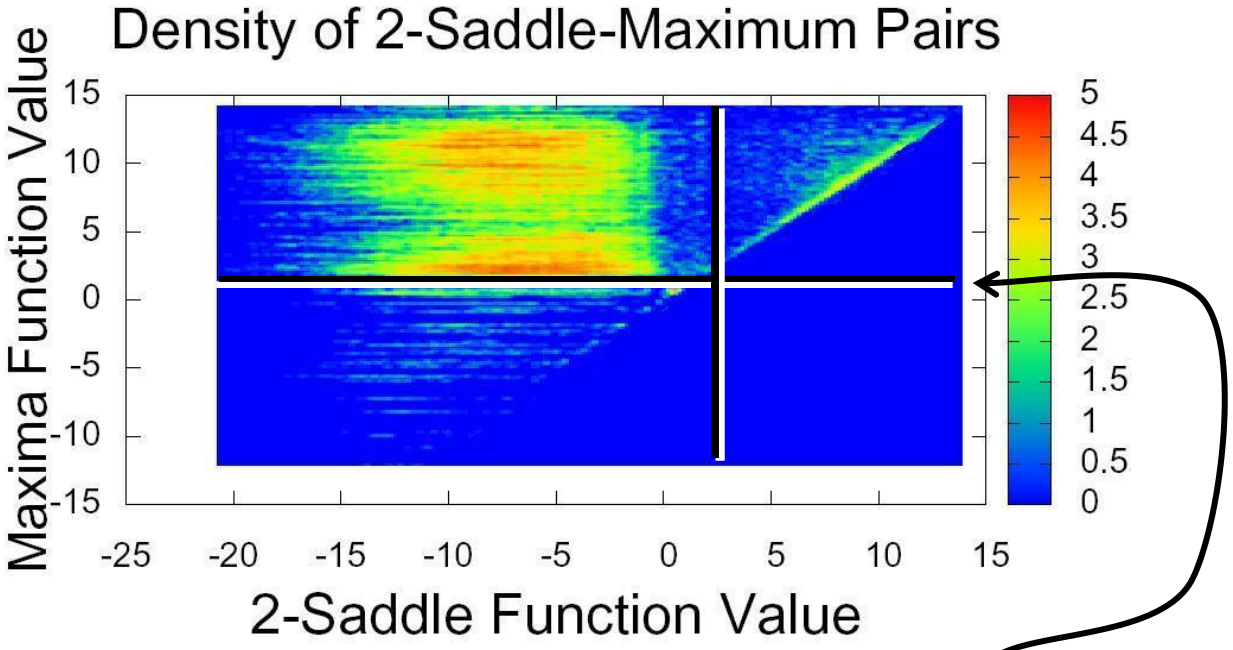
Need to Find Proper Threshold Values and Characterize the Stability of the Solution

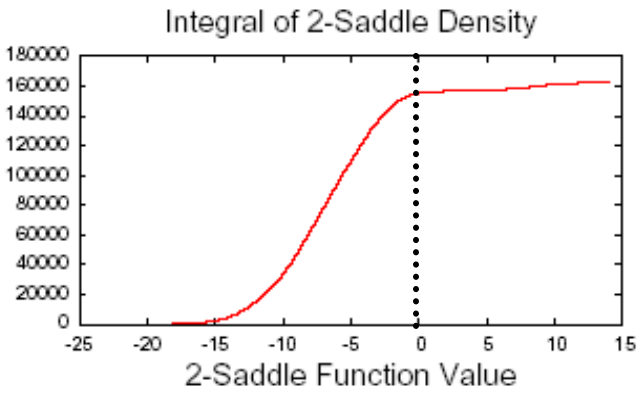


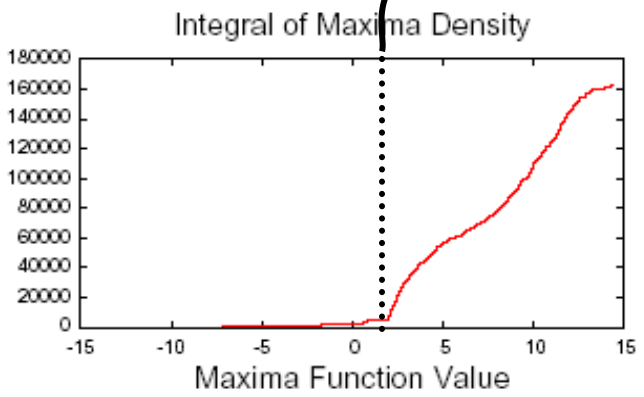




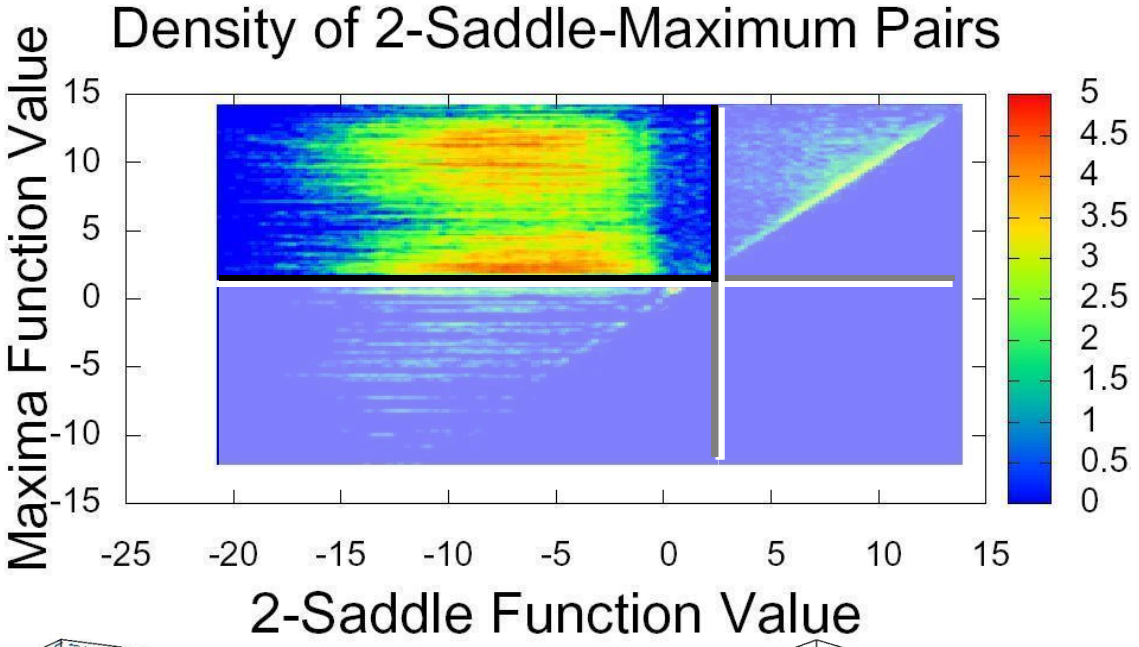
Need to Find Proper Threshold Values and Characterize the Stability of the Solution

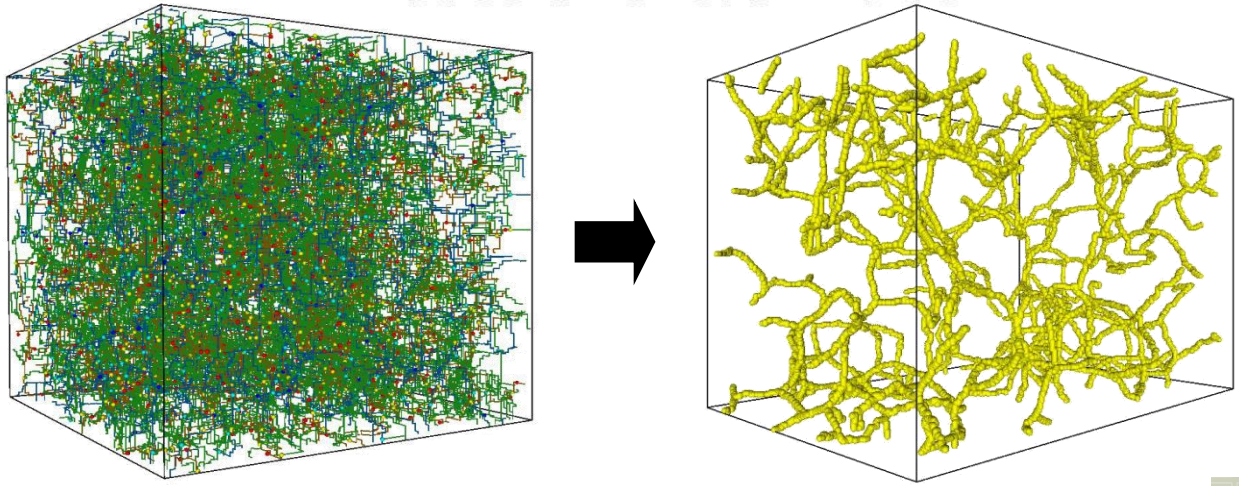




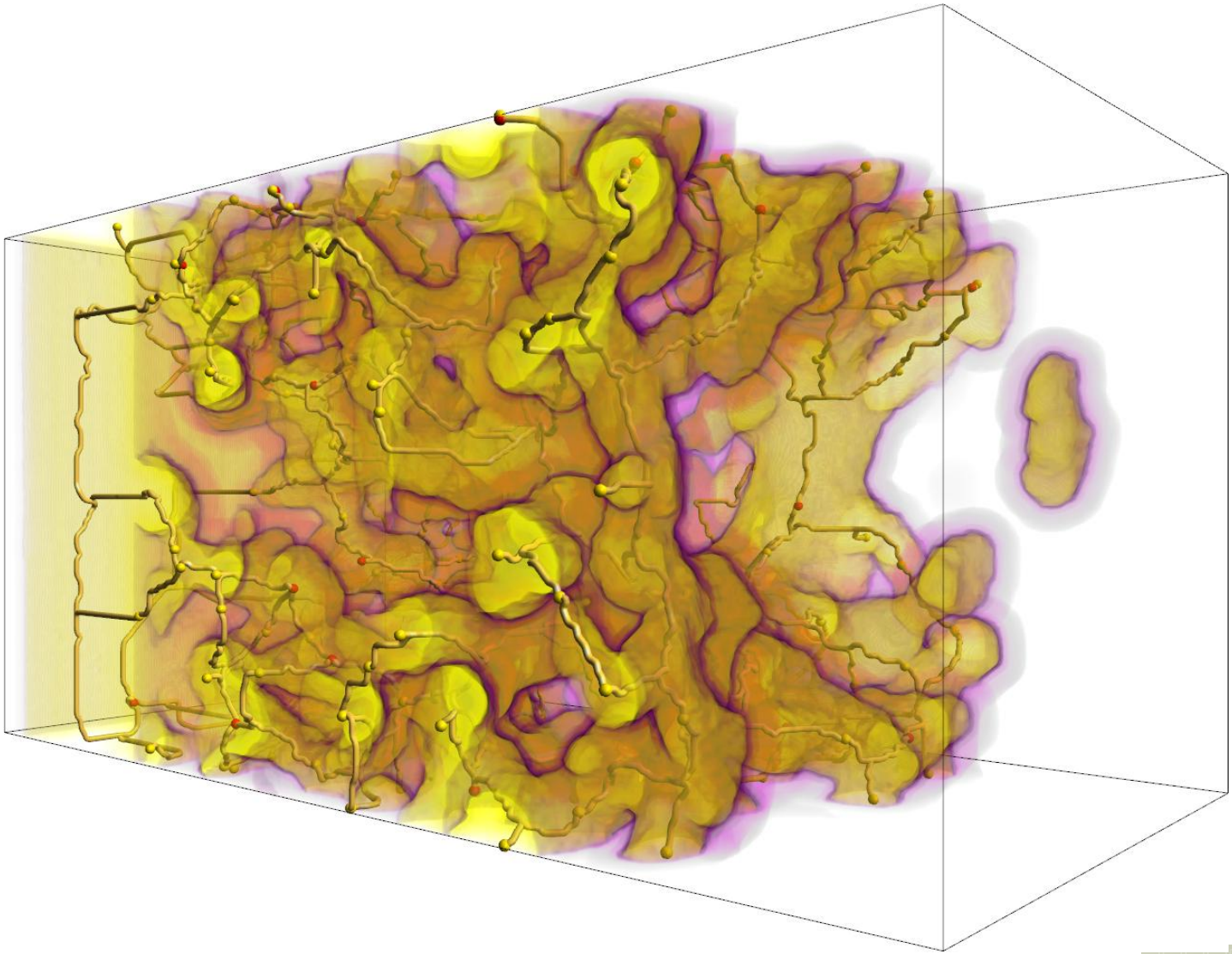


We Obtain a Robust Reconstruction of the Filament Structures in the Material



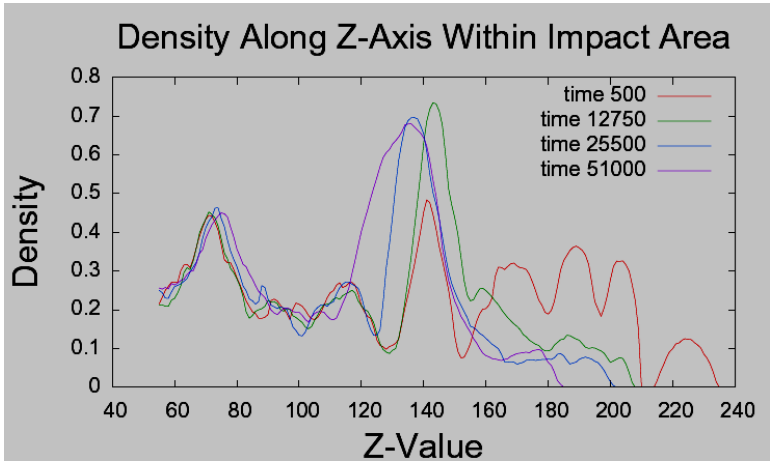


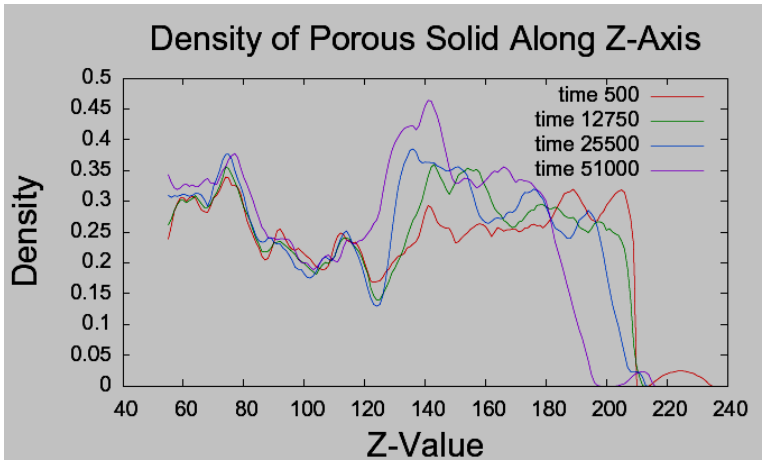
Demo Porous Medium



The Extracted Structures Allow to Quantify the Change in Porosity of the Material

Density profiles

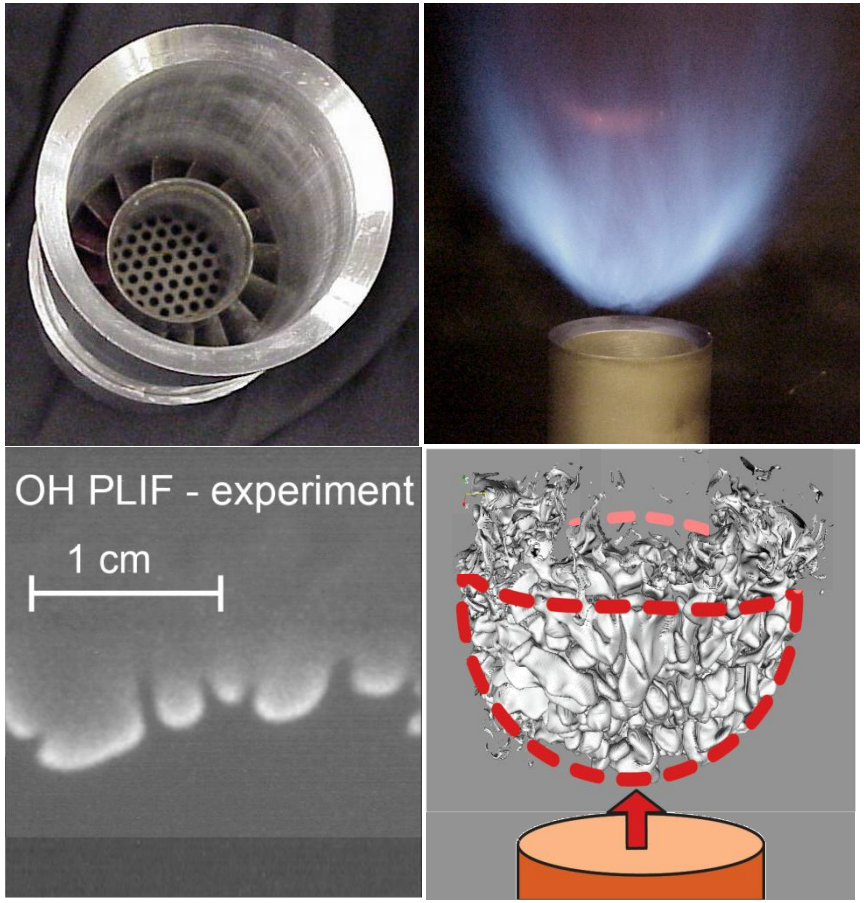




Decay in porosity of the material

Metric	t=500	t=12750	t=25500	t=51000
# Cycles	762	340	372	256
Total Length	34756	24316	23798	18912

Understanding Turbulence for Low Emission, High Efficiency Combustion

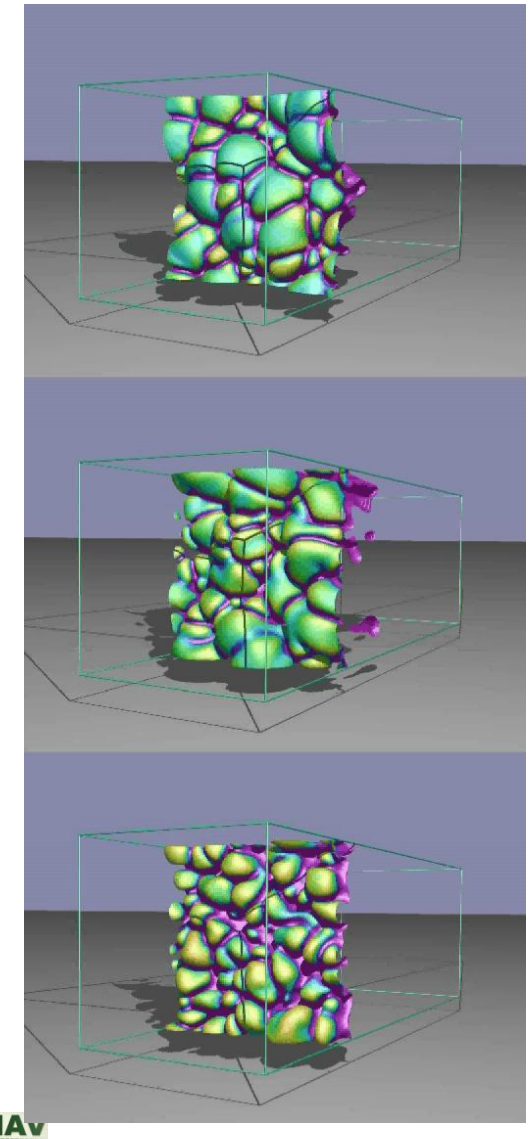


Experiment Simulation

- Lean premixed H_2 flames
- Low Swirl Combustion (LSC) Burners
- Low pollution in energy production
- High Efficiency in fuel consumption
- Scalable from residential to industrial use
- Each variable 3.9-4.5 TB

1" burner (5 kW, 17 KBtu/hr) 28" burner (44 MW, 150 MBtu/hr)

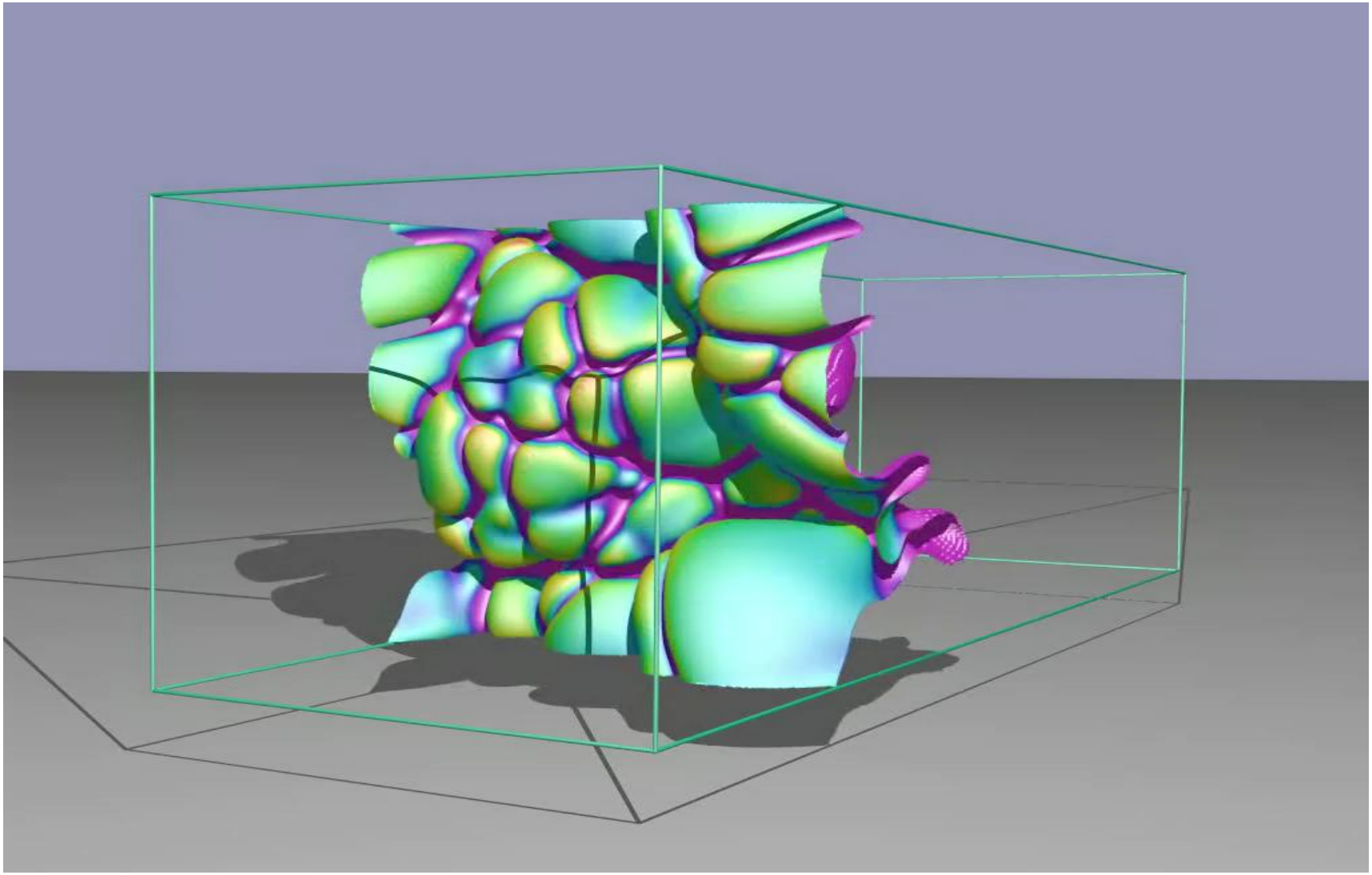
We Take on the Challenge of Developing a Quantitative Analysis Detecting Turbulence



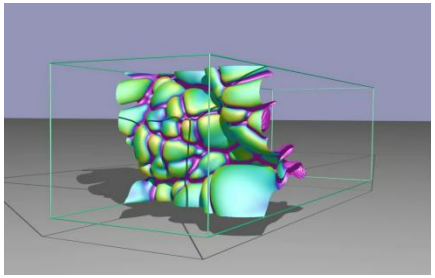
Understanding combustion processes over a broad range of burning conditions is an important problem for designing engines and power plants.

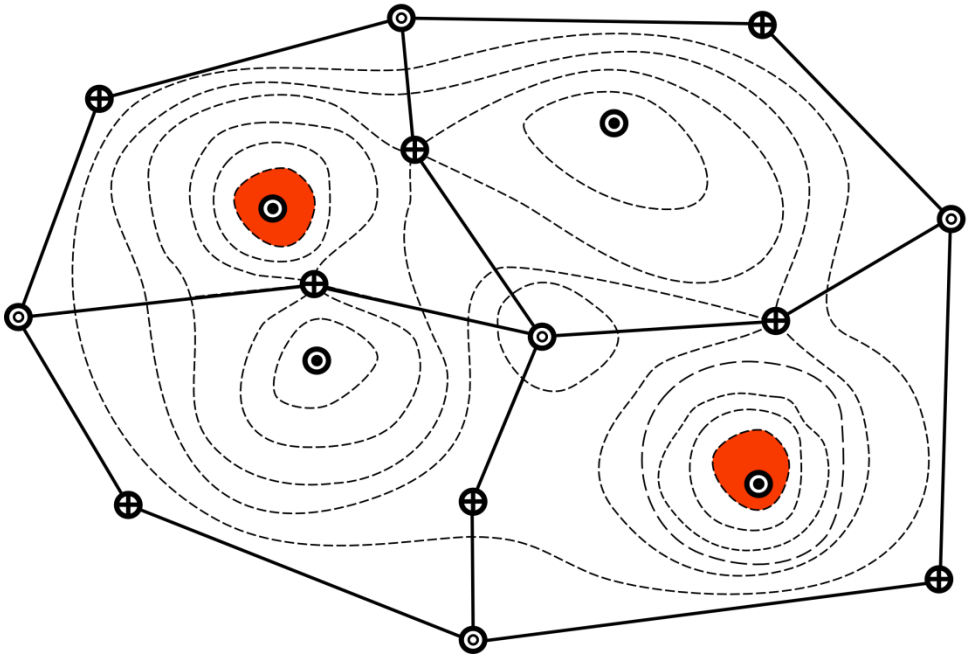
- **Simulation with AMR mesh.**
- **Simulations of lean premixed hydrogen flames with three degrees of turbulence.**
- **Can we identify precisely and track in time burning regions?**
- **Can we discriminate the degree of turbulence from a quantitative analysis?**

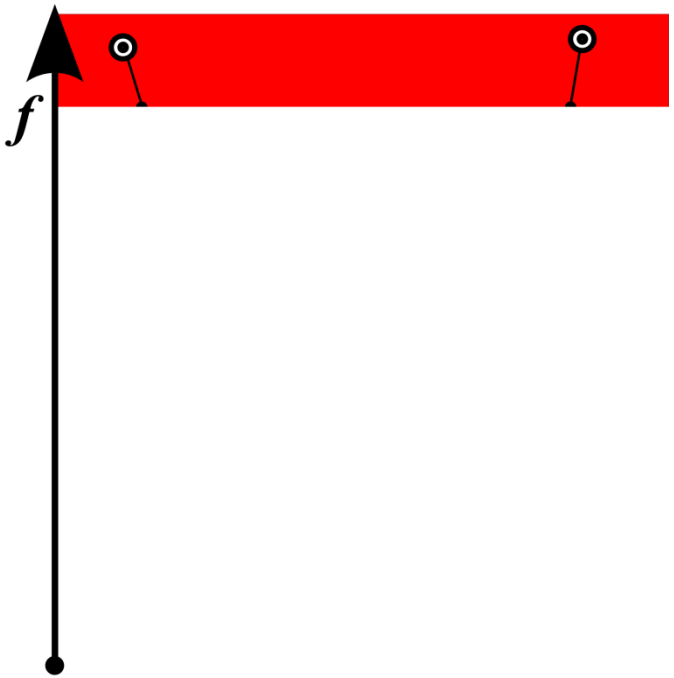
We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



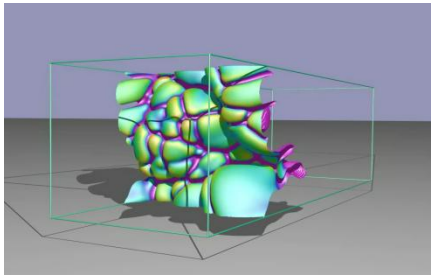


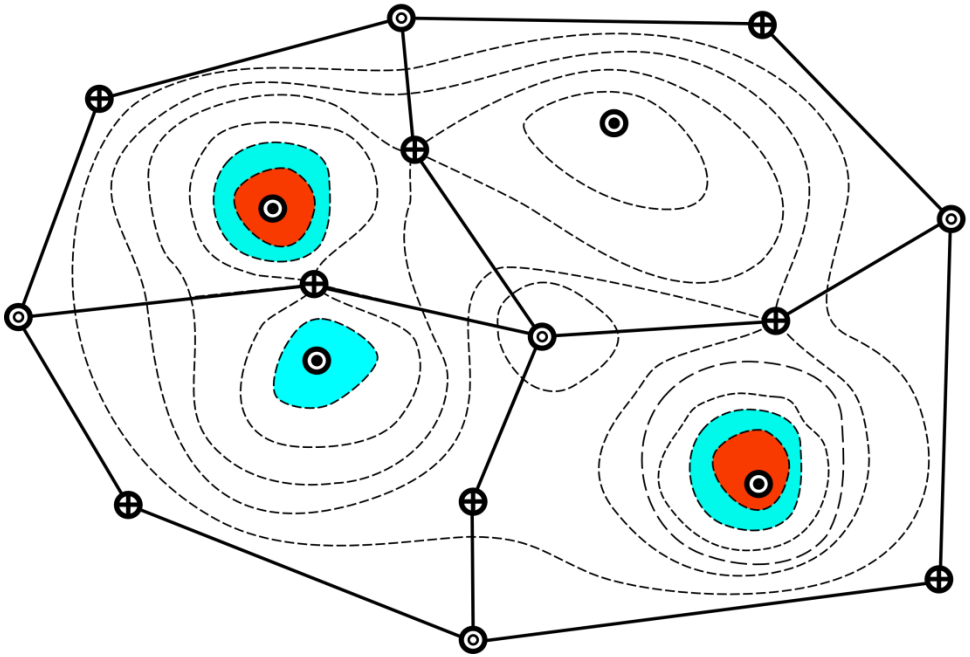


domain

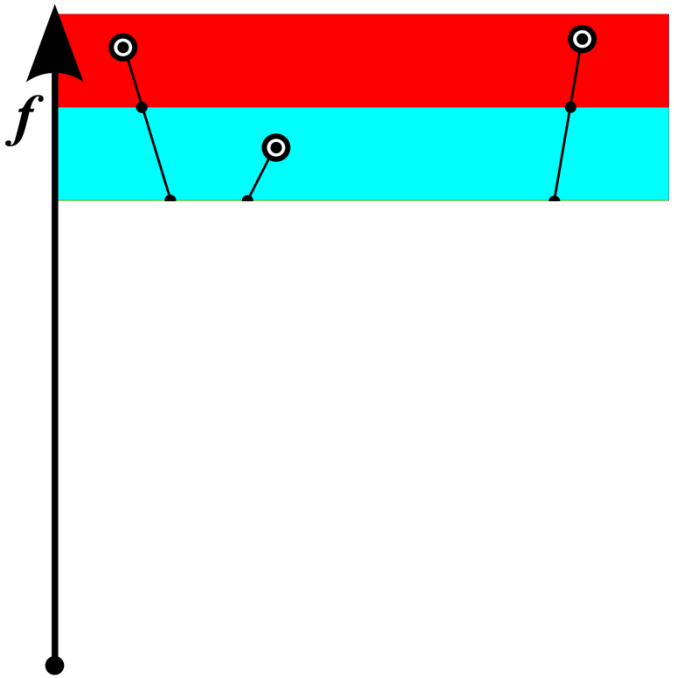
hierarchy

We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



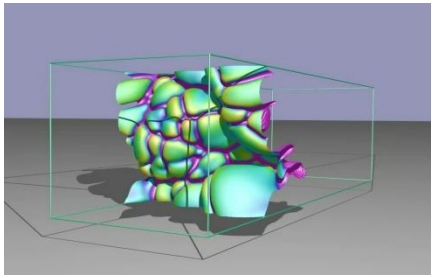


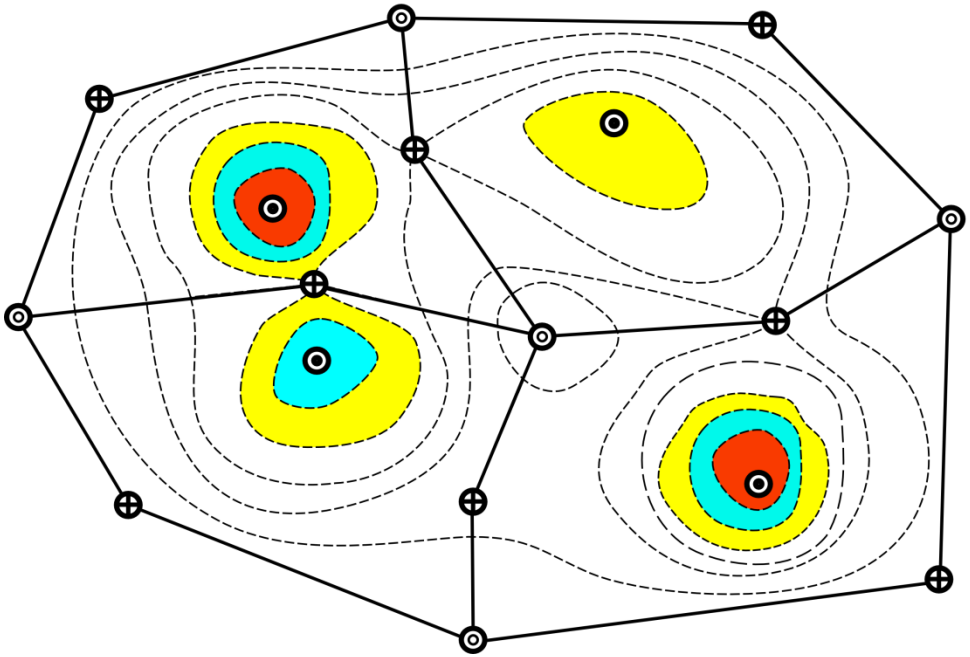
domain



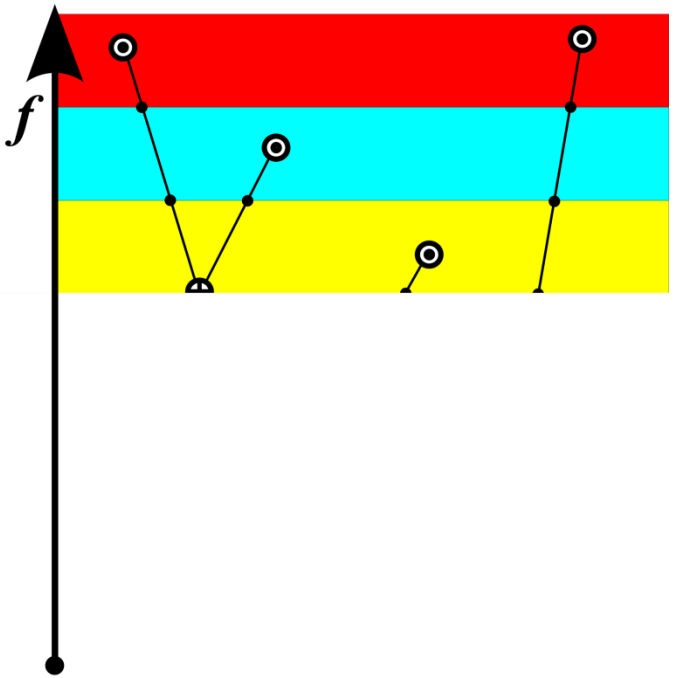
hierarchy

We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



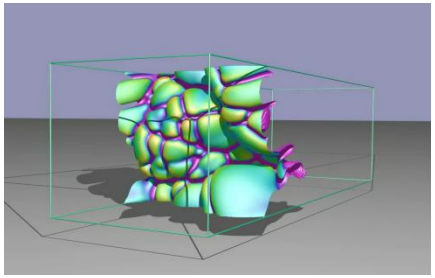


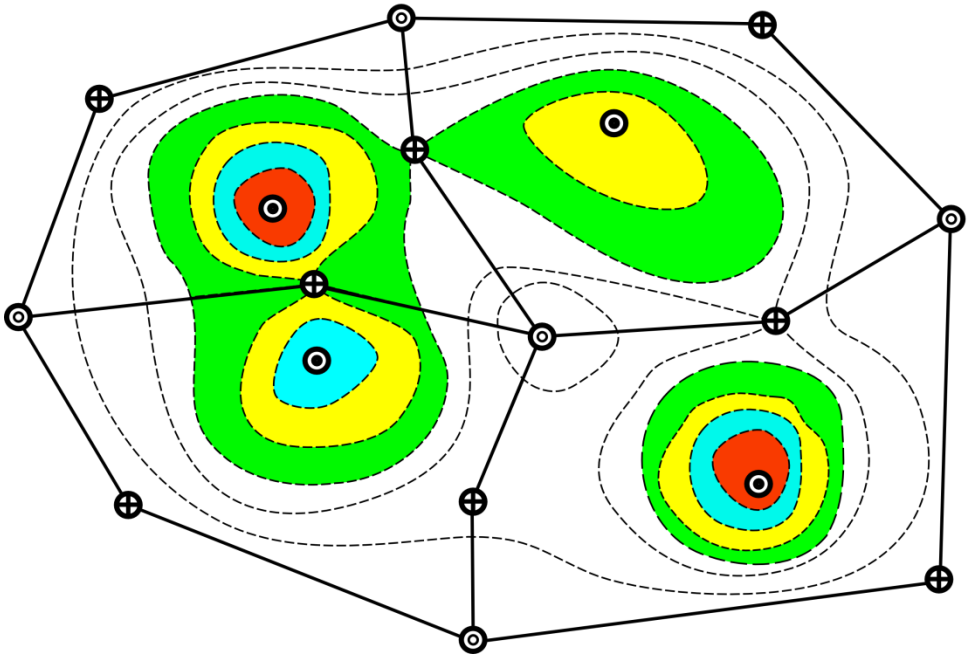
domain



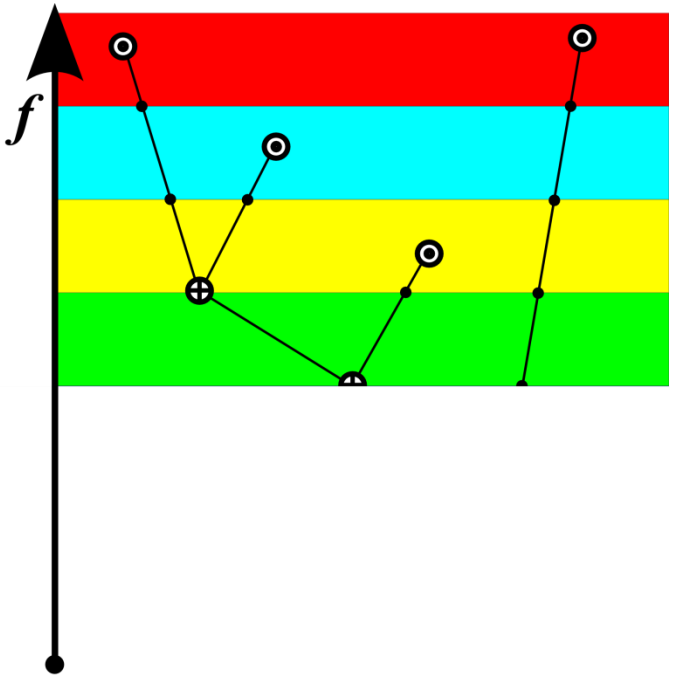
hierarchy

We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



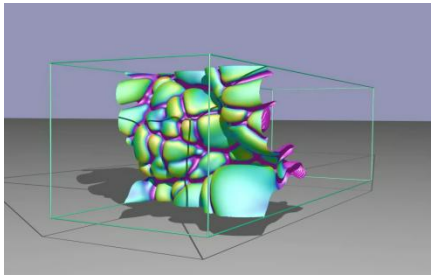


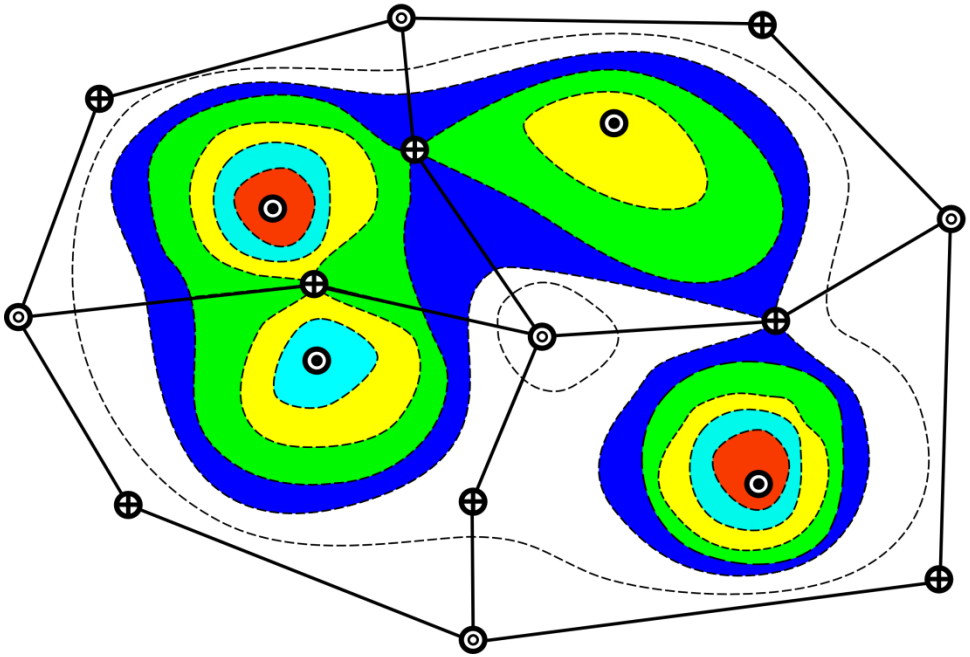
domain

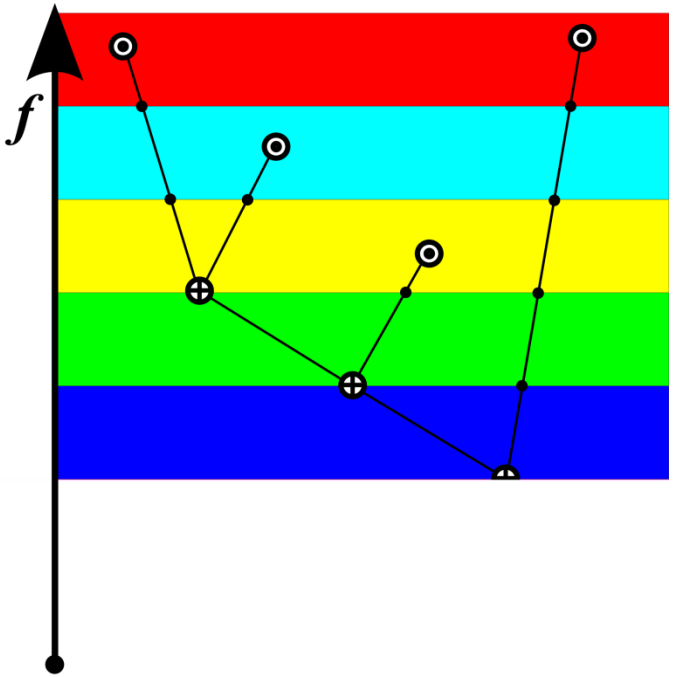


hierarchy

We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



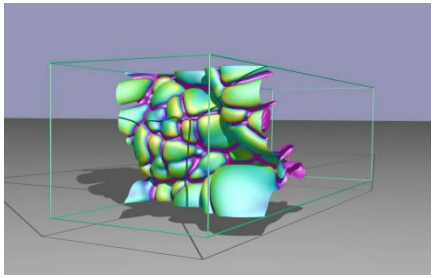


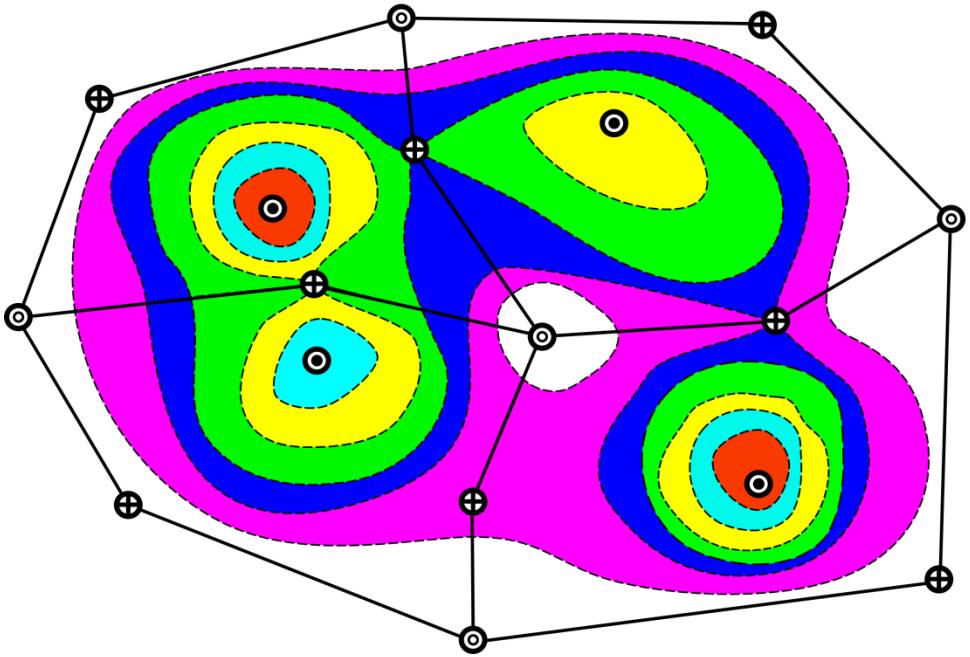


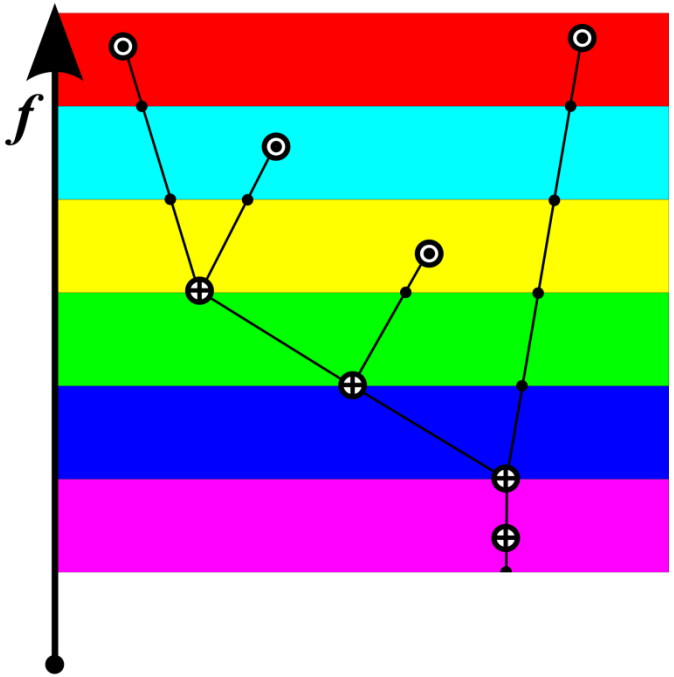
domain

hierarchy

We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



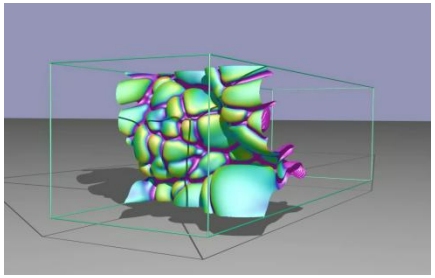


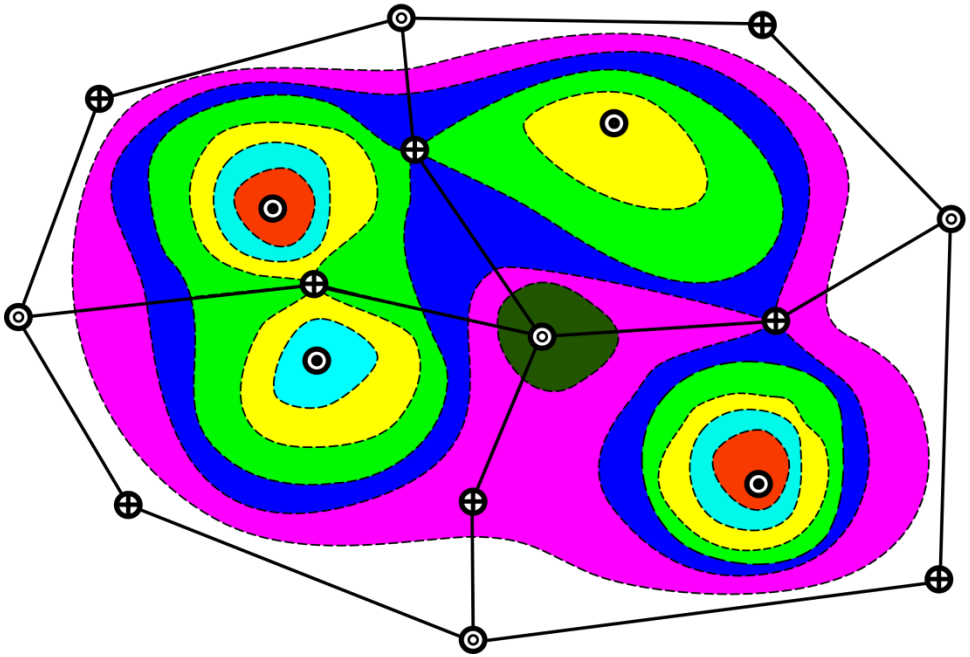


domain

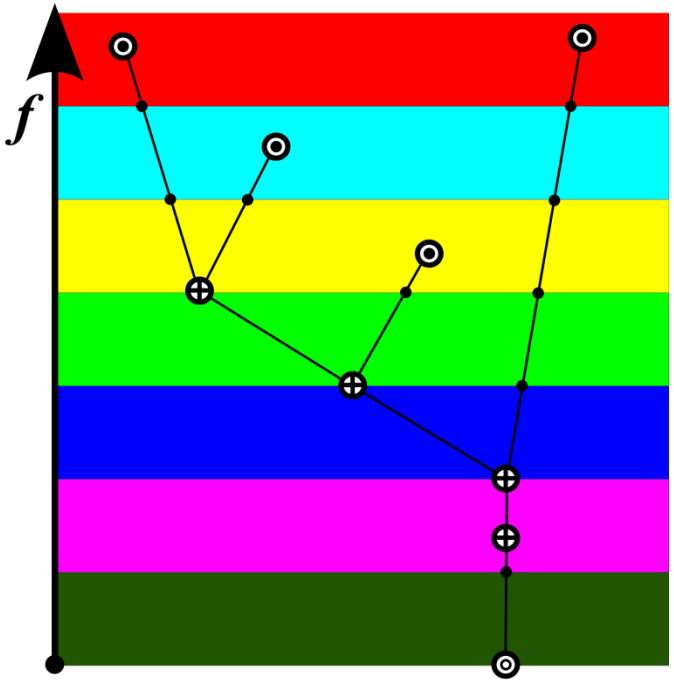
hierarchy

We Build a Reduced Topological Model of H2 Consumption on an Isotermal Surfaces



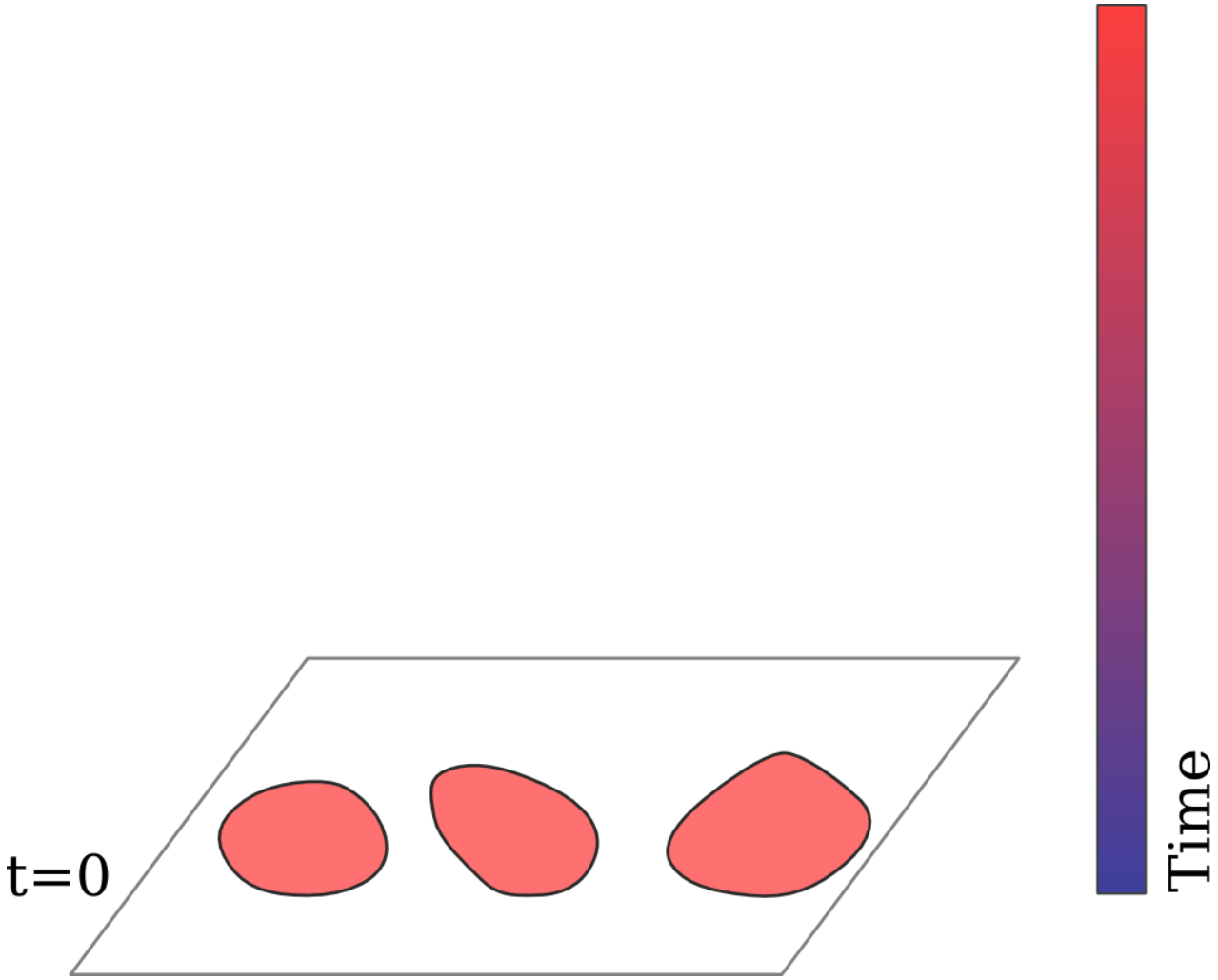


domain

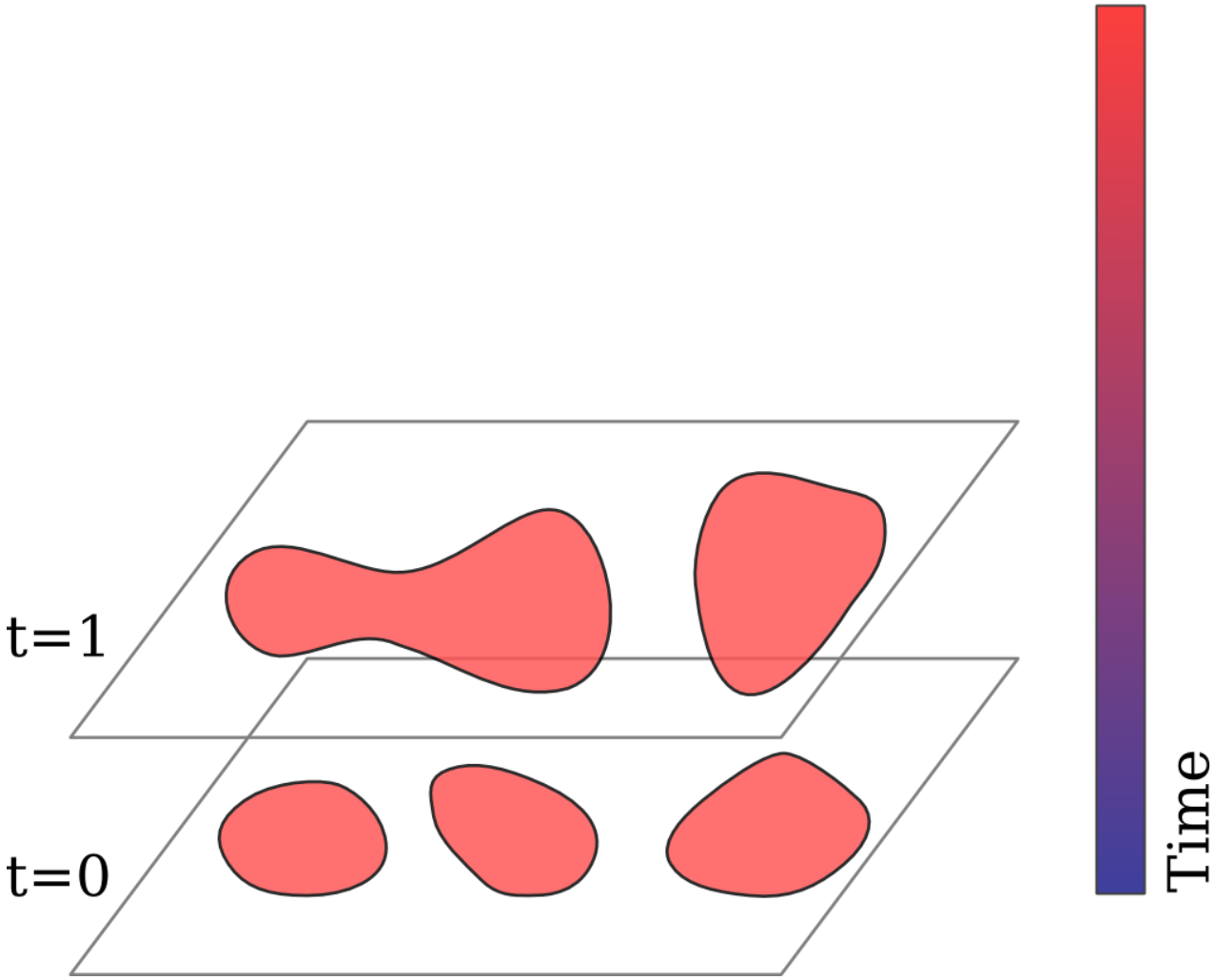


hierarchy

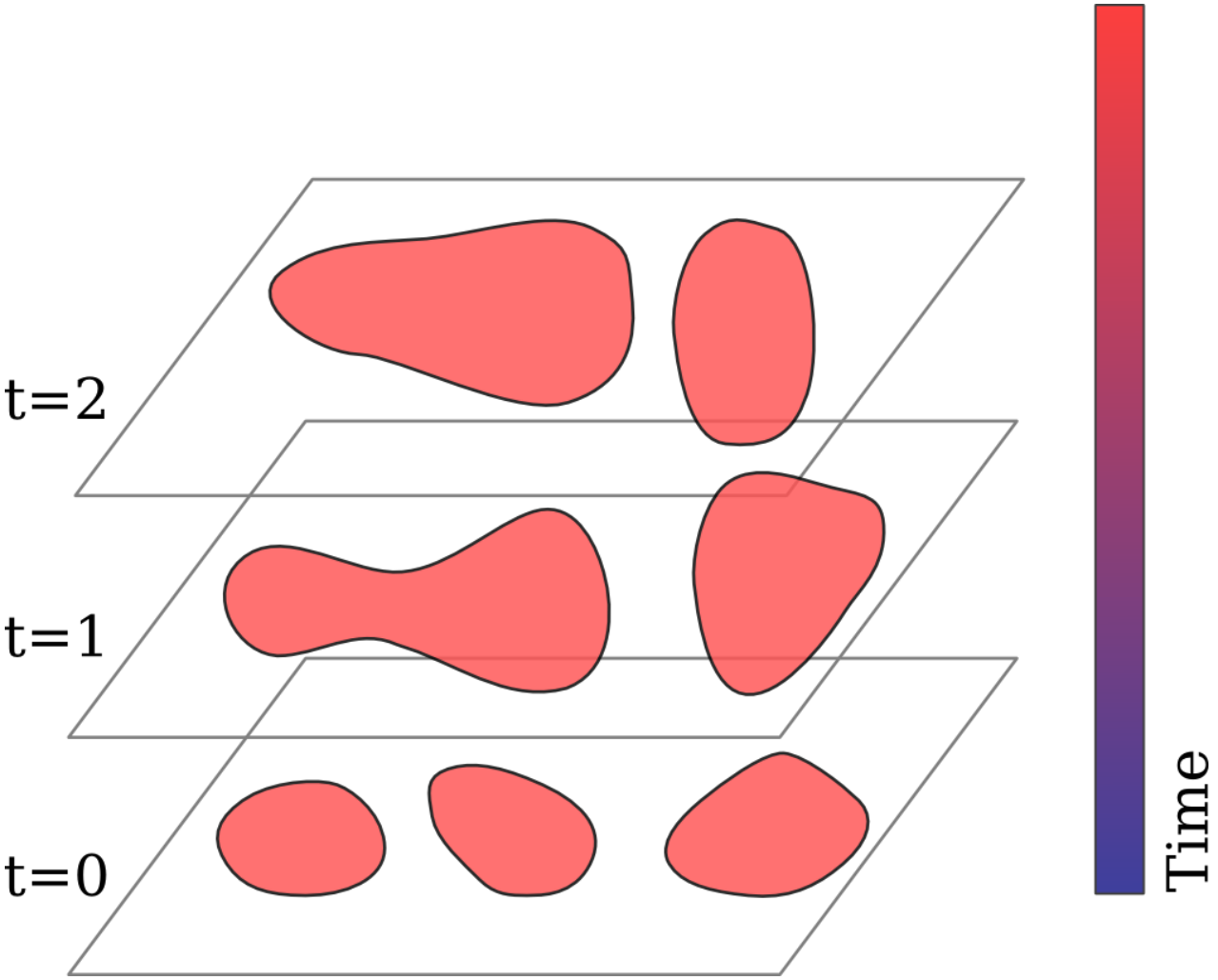
We track in time by interpolation in 4D and contraction to a Reeb Graph



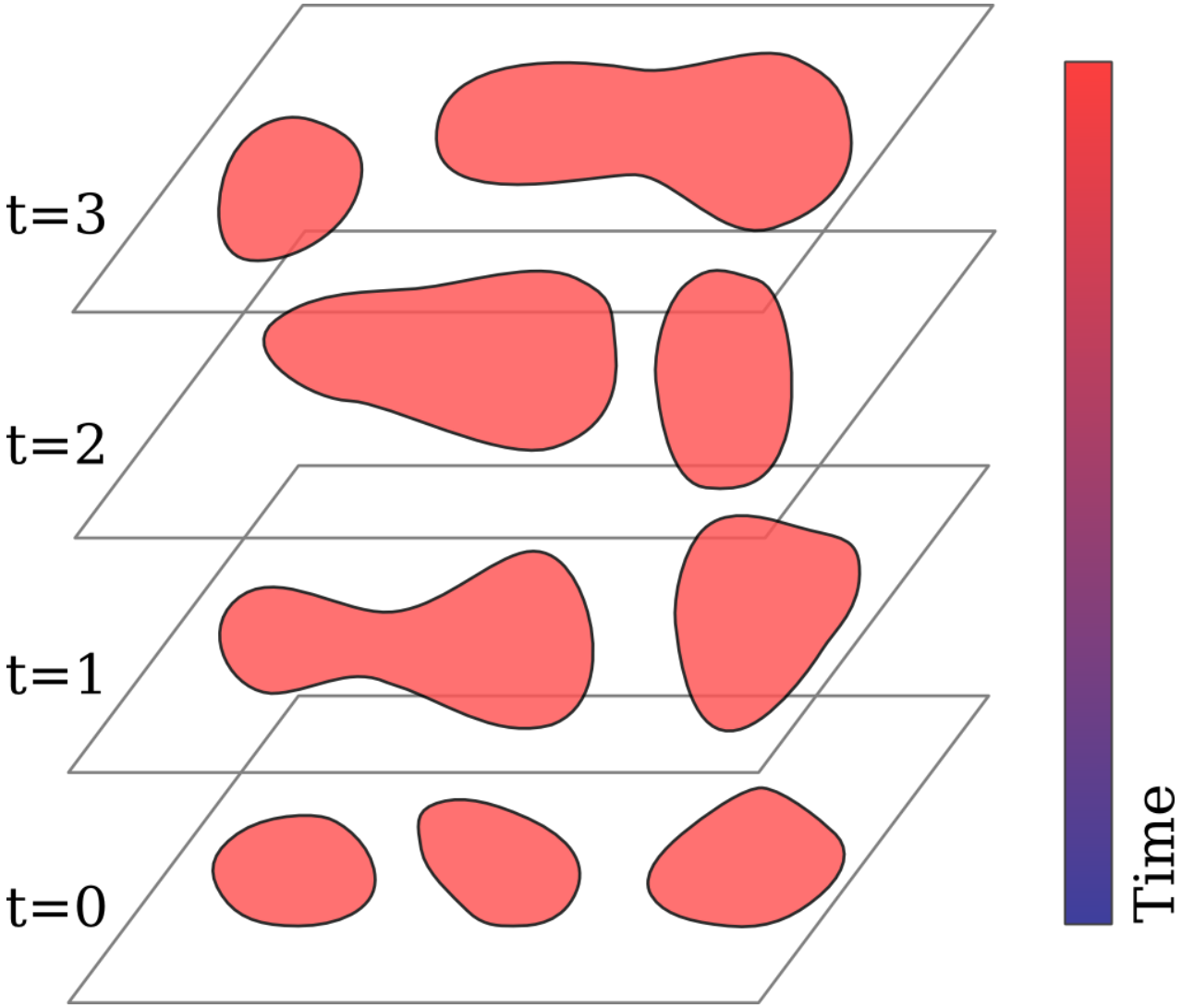
We track in time by interpolation in 4D and contraction to a Reeb Graph



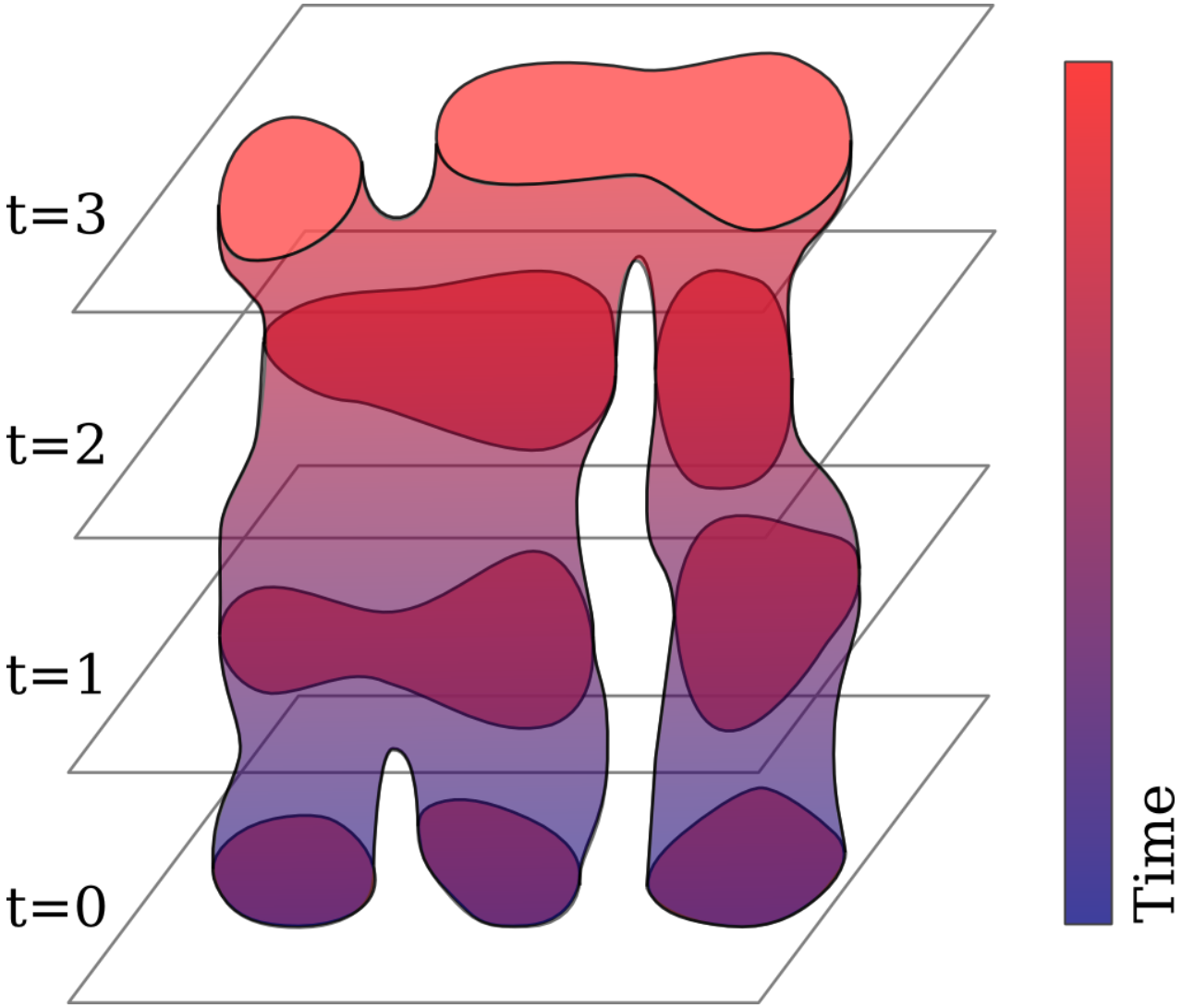
We track in time by interpolation in 4D and contraction to a Reeb Graph



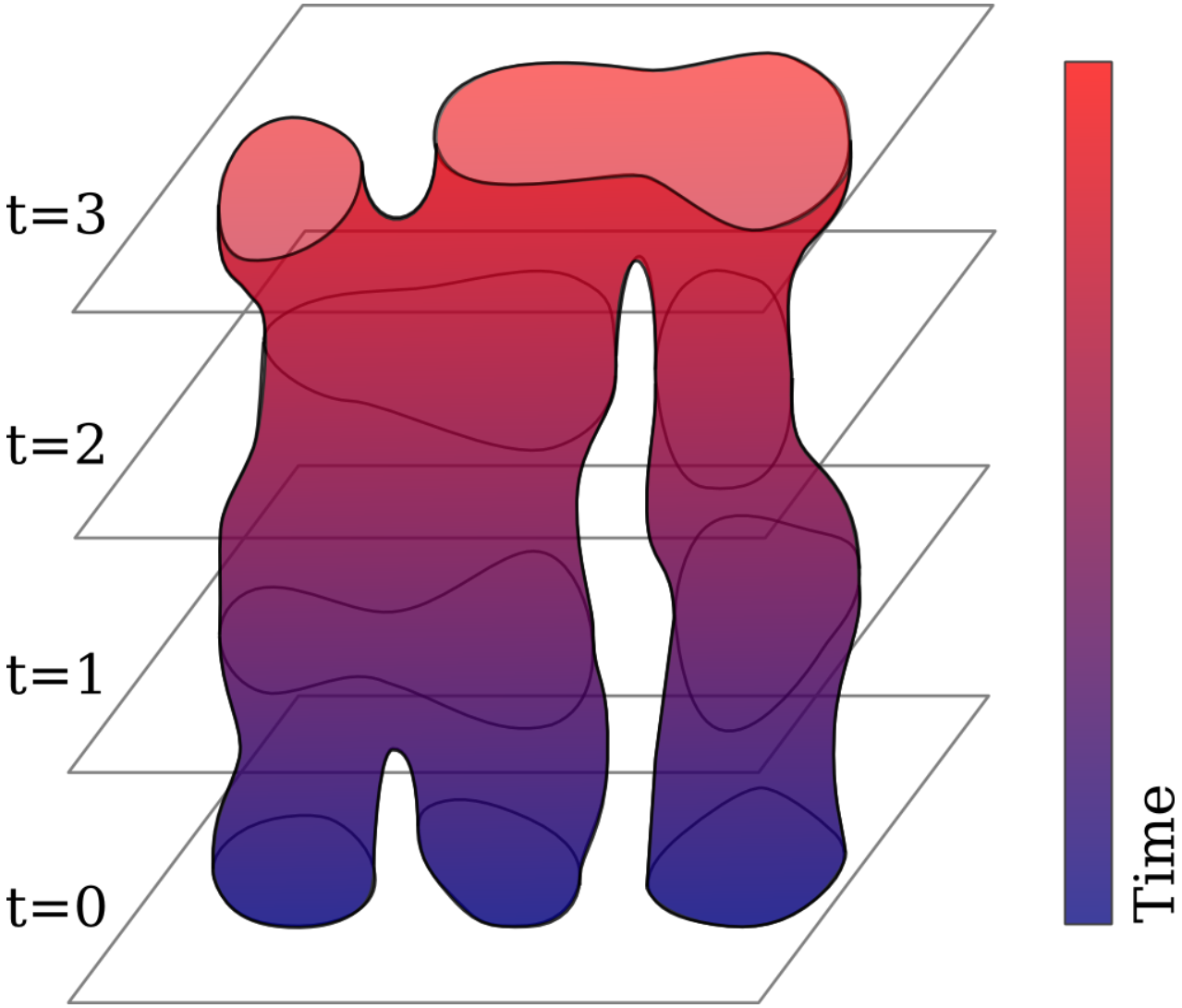
We track in time by interpolation in 4D and contraction to a Reeb Graph



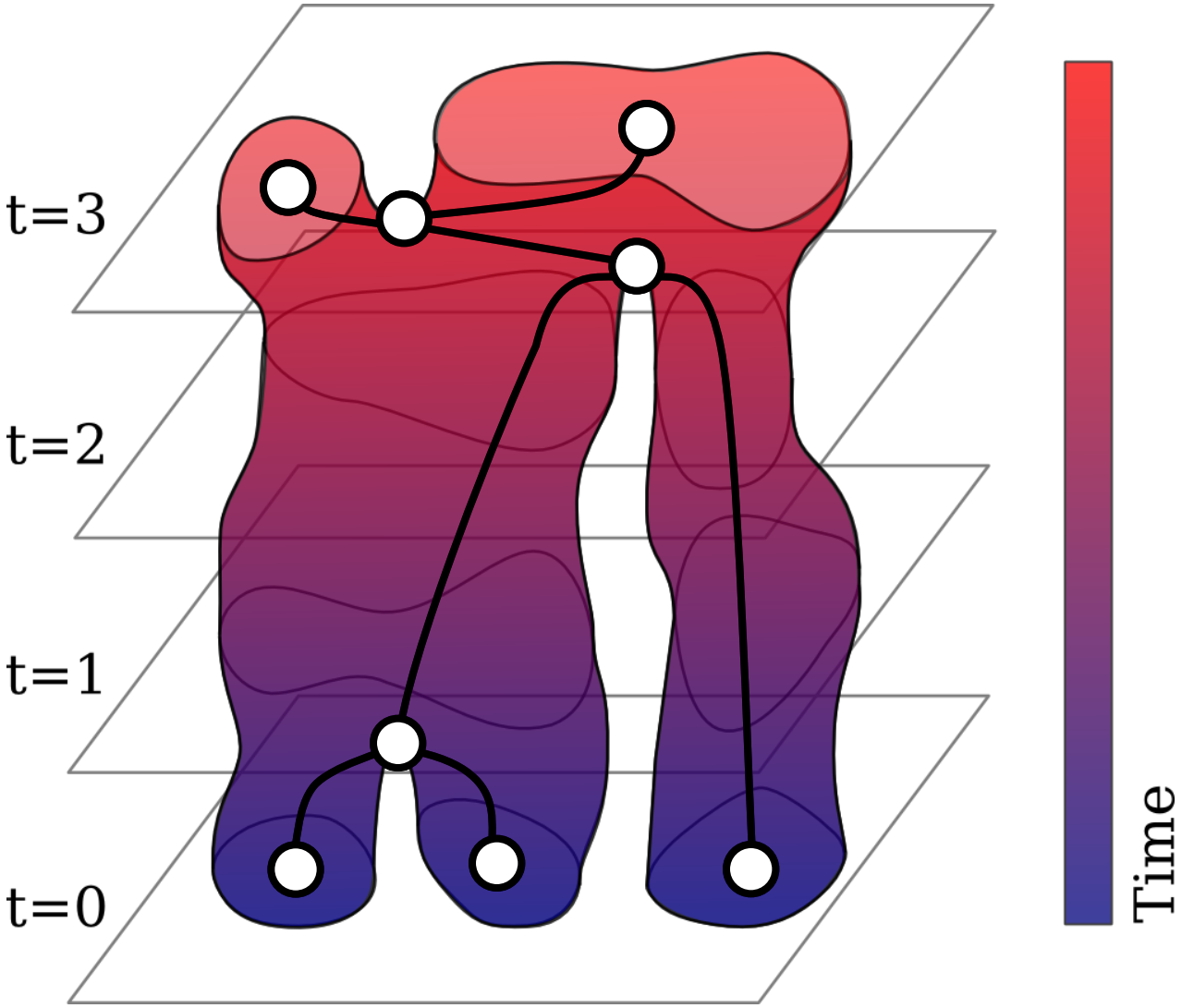
We track in time by interpolation in 4D and contraction to a Reeb Graph



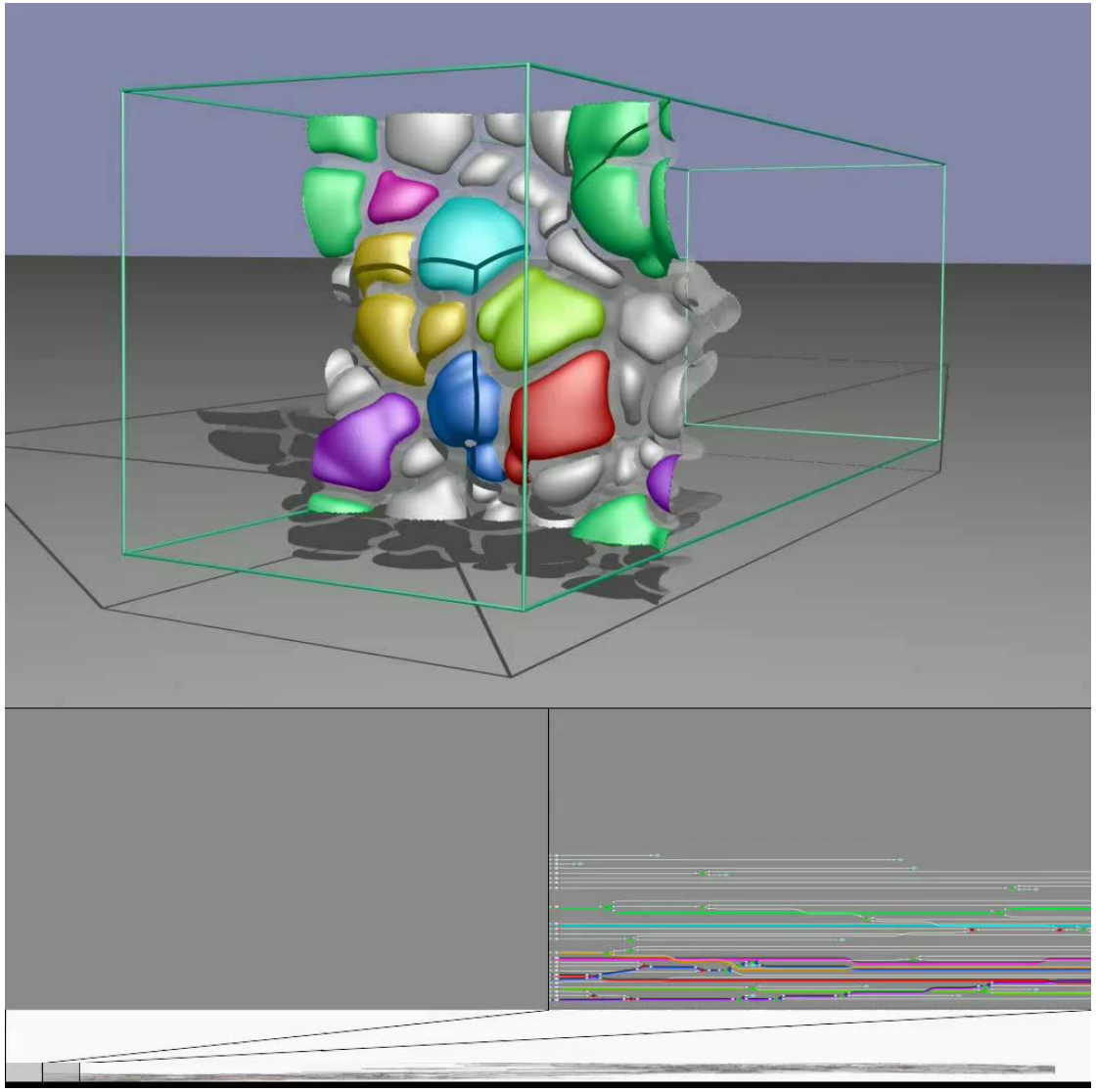
We track in time by interpolation in 4D and contraction to a Reeb Graph



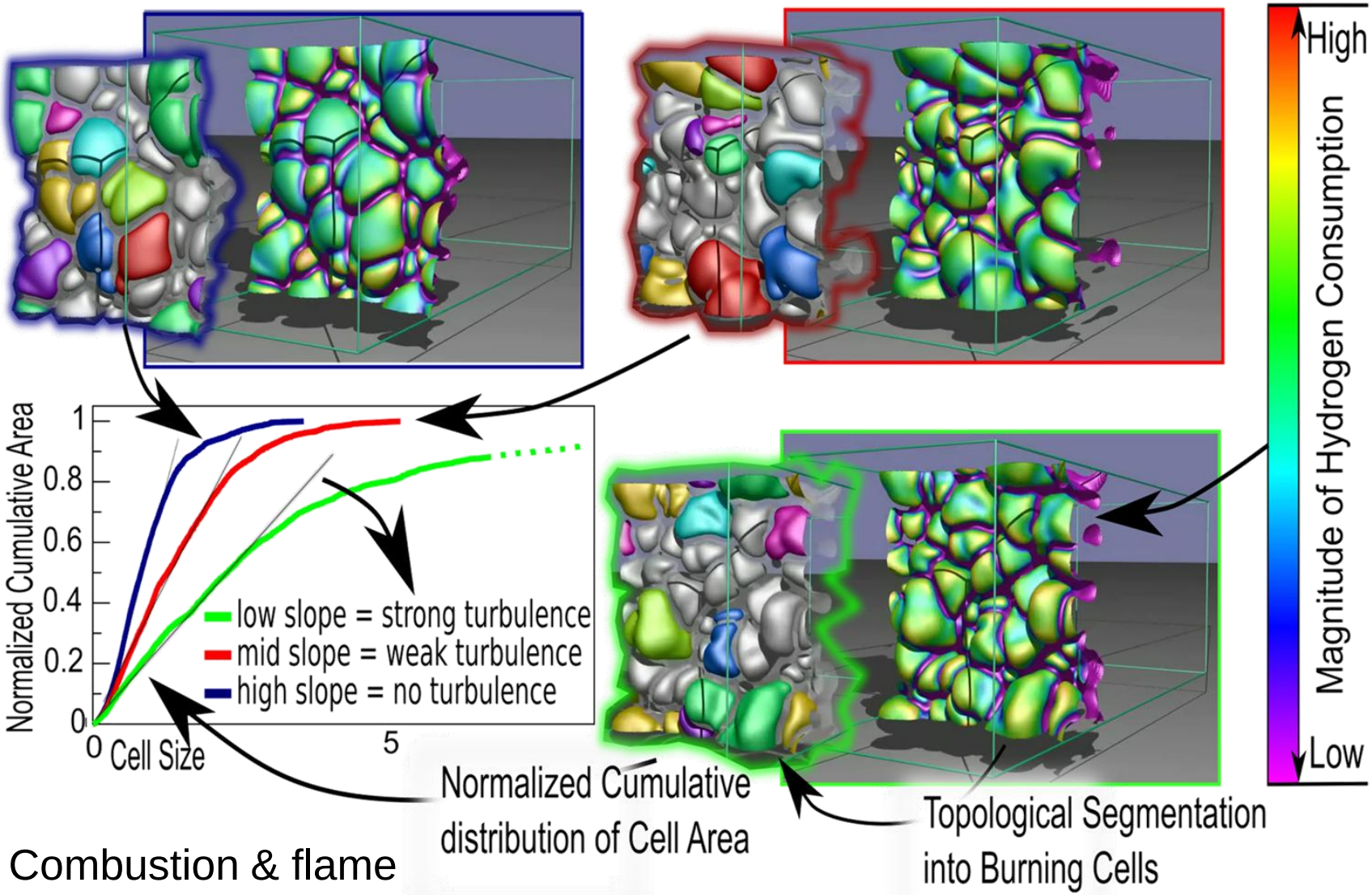
We track in time by interpolation in 4D and contraction to a Reeb Graph



Each Set of Parameters Results in a Robust Segmentation and Tracking of Burning Cells



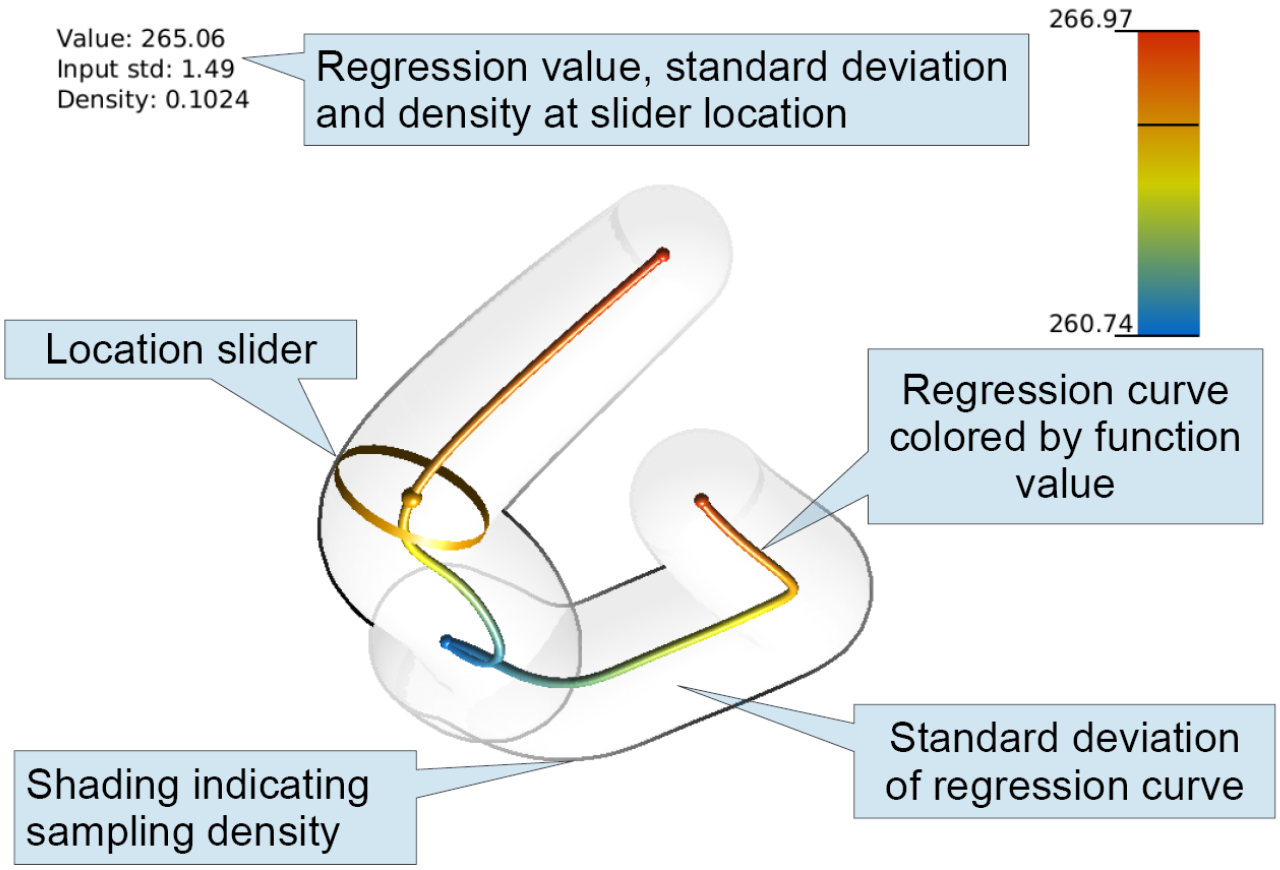
Topological Segmentation Allows to Quantify Turbulence as Slope of the Area Distributions



- Combustion & flame

Exploration of High Dimensional Functions for Sensitivity Analysis

Integrated presentation of statistics and topology



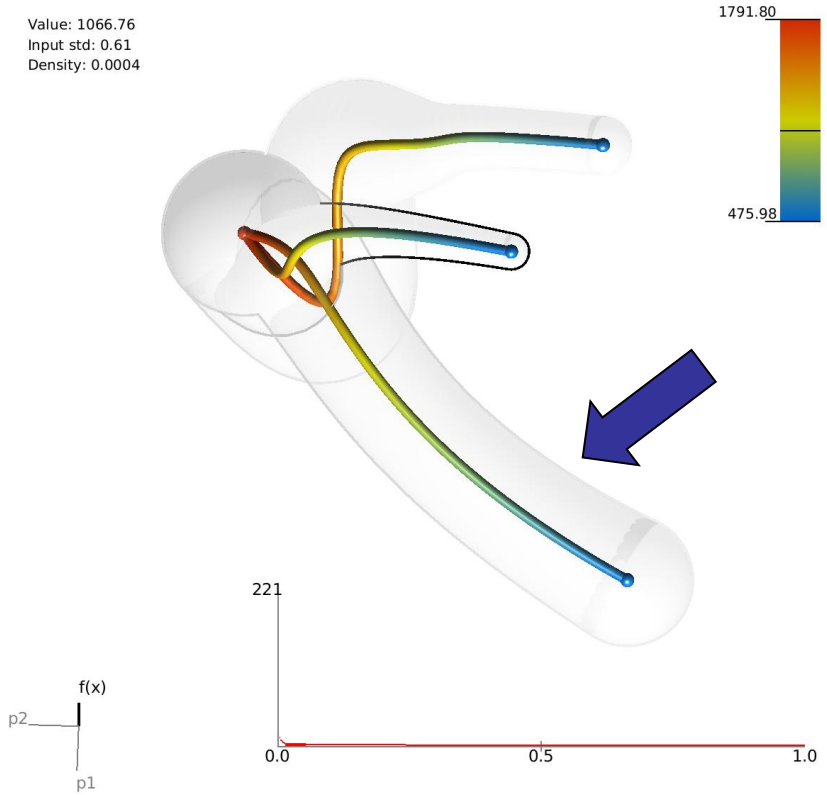
Analysis of Combustion Simulations

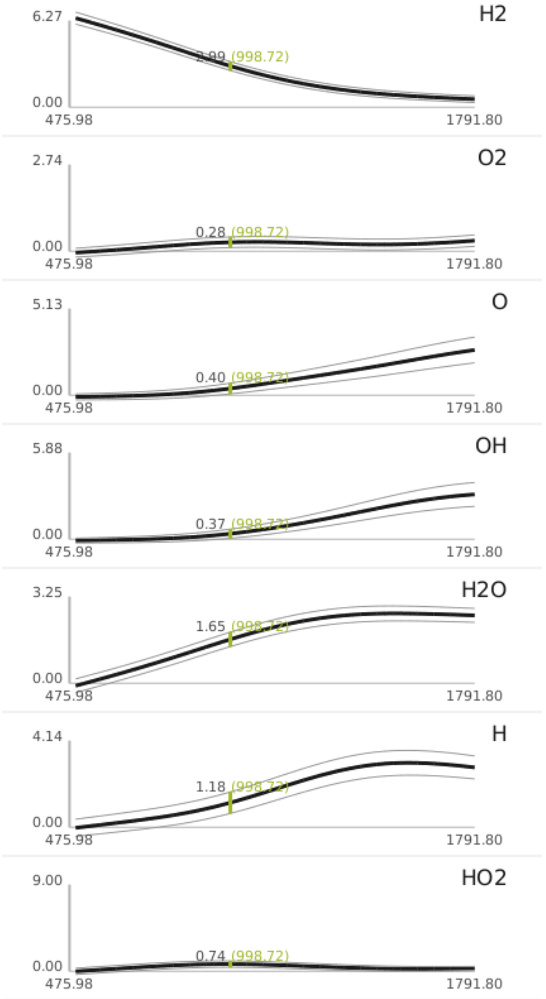
Combustion Simulation of Jet CO/H2-Air Flames

Input: Composition of 10 chemical species

Output: Temperature

The Framework Allows Detailed Visualization and Analysis of High Dimensional Functions

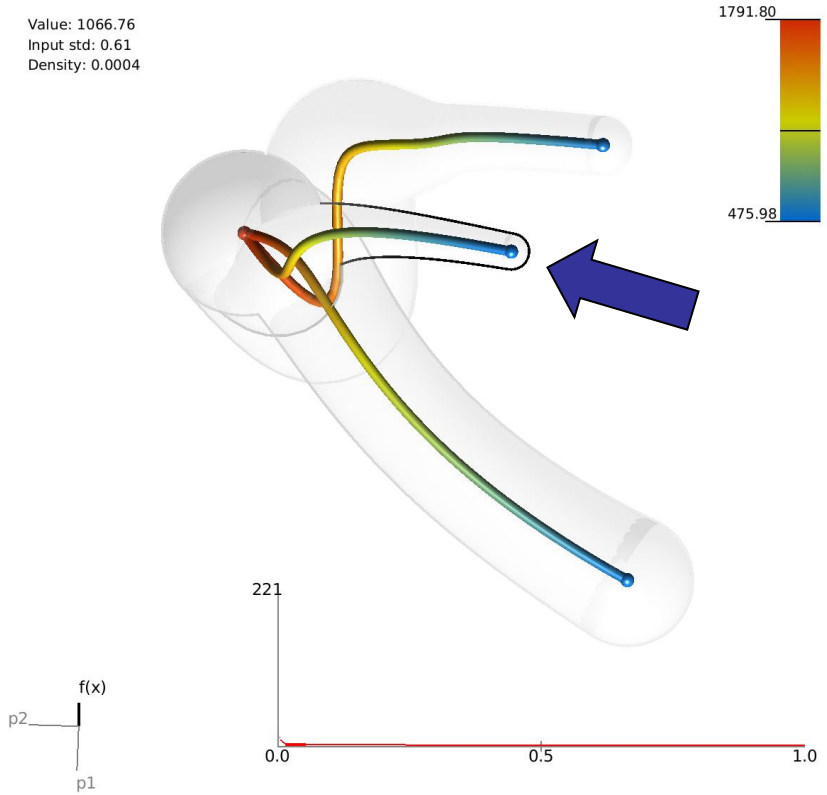


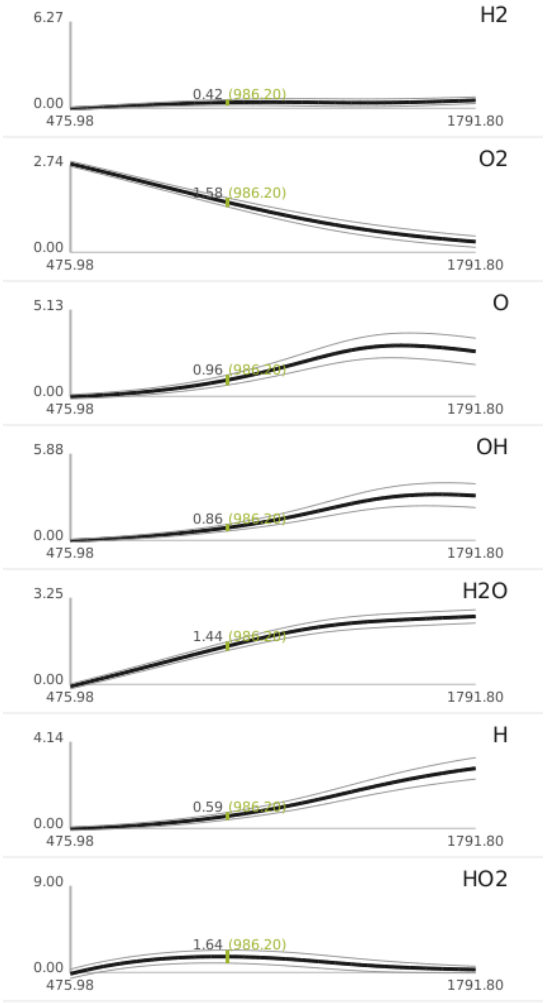


Pure fuel

10 dimensional data set describing the heat release wrt. to various chemical species in a combustion simulation

The Framework Allows Detailed Visualization and Analysis of High Dimensional Functions

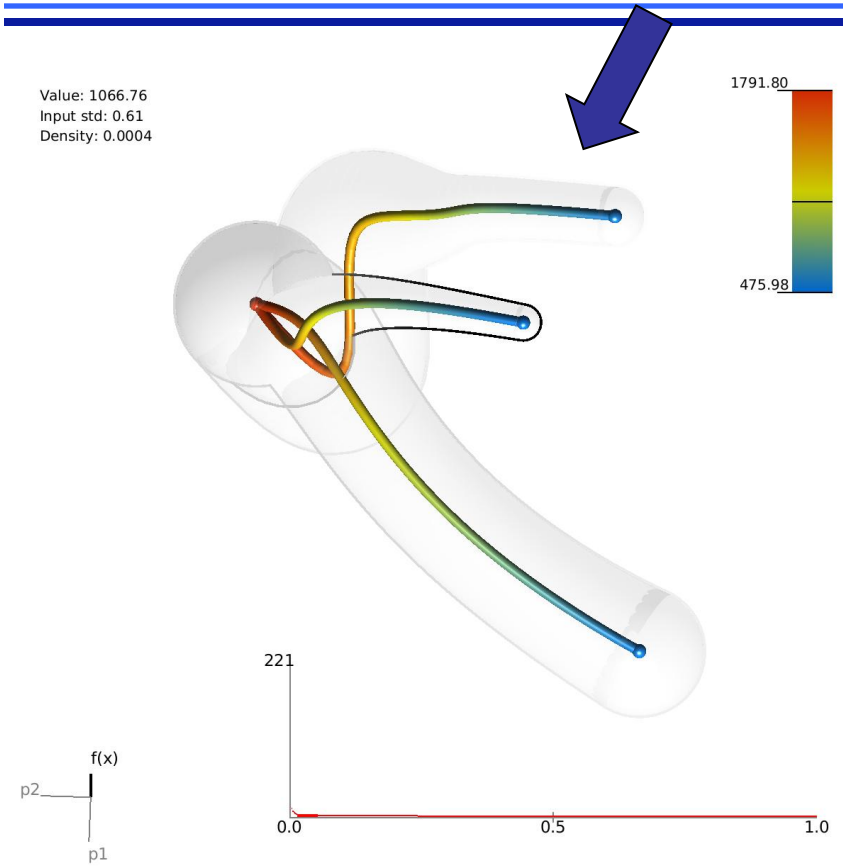


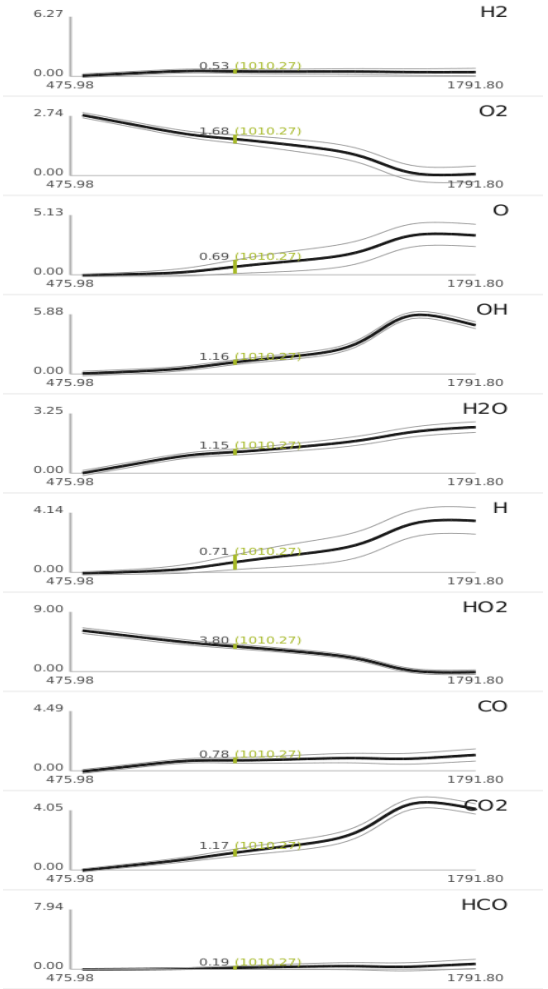


Pure oxidizer

10 dimensional data set describing the heat release wrt. to various chemical species in a combustion simulation

The Framework Allows Detailed Visualization and Analysis of High Dimensional Functions





Local extinction

10 dimensional data set describing the heat release wrt. to various chemical species in a combustion simulation

Combustion Simulation of Jet CO/H2-Air Flames

Input: Composition of 10 chemical species

Output: Temperature

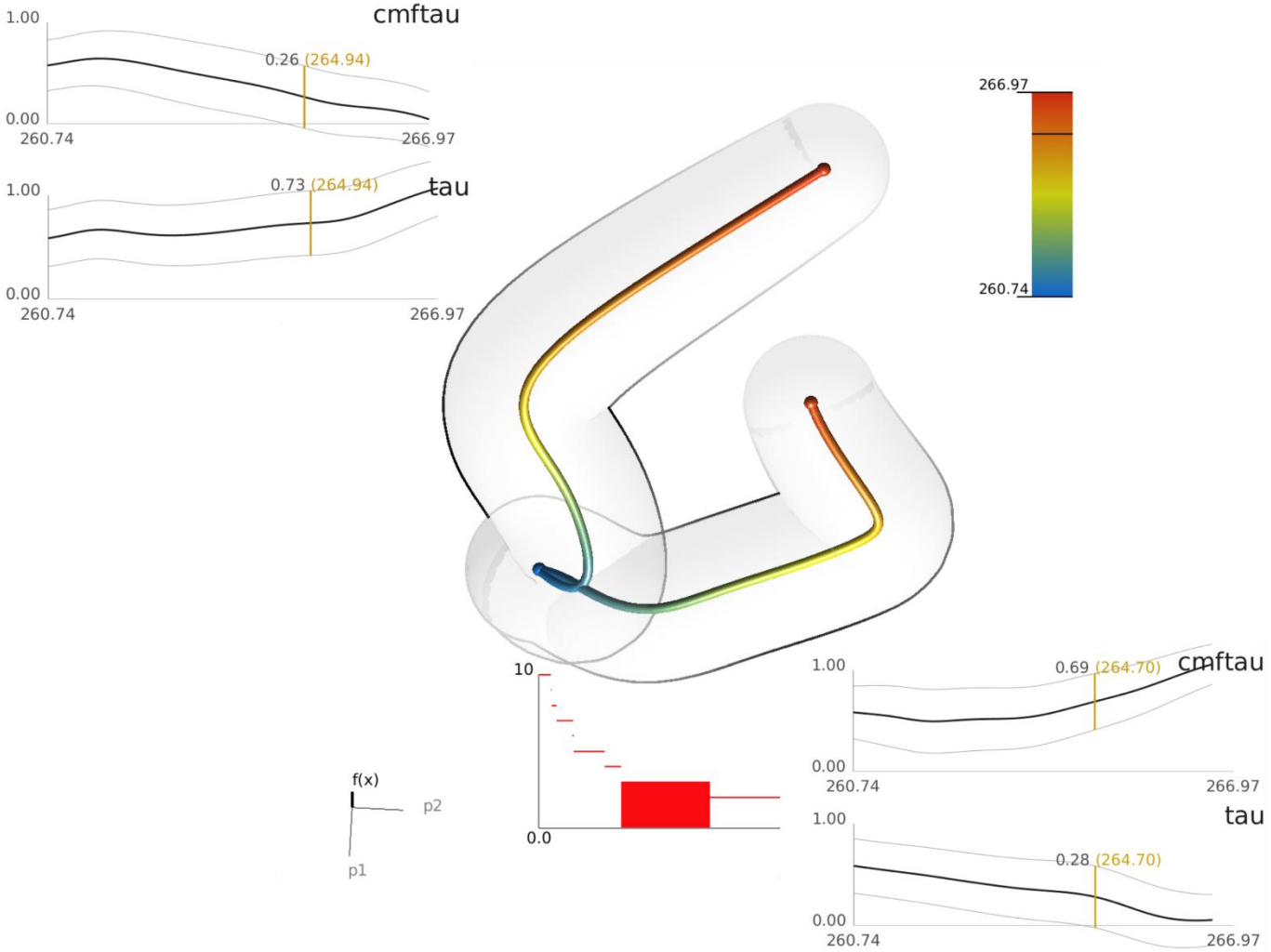
Analysis of Climate Data

Community Atmosphere Climate Model

Input: 21 parameter settings

Output: Net long wave flux (thermal radiation)

The Framework Reveals Relationship Between Convection and Global Long Wave Flux



Community Atmosphere Climate Model

Input: 21 parameter settings

Output: Net long wave flux (thermal radiation)

Data Analysis and Visualization Center is a Catalyst for a Virtuous Cycle of Collaborative Activities

- **Tight cycle of :**
  - **basic research,**
  - **software deployment**
  - **user support**
- **Coordination among many projects:**
  - **unified techniques for several applications**
- **Strong University-Lab-Industry collaboration**
- **Focused technical approach:**
  - **performance tools for fast data access**
  - **general purpose data exploration**
  - **error bounded quantitative analysis**
  - **feature extraction and tracking**
- **Interdisciplinary collaboration with domain scientists (from math to physics):**
  - **motivating the work**
  - **formal theoretical approaches**
  - **feedback to specific disciplines**