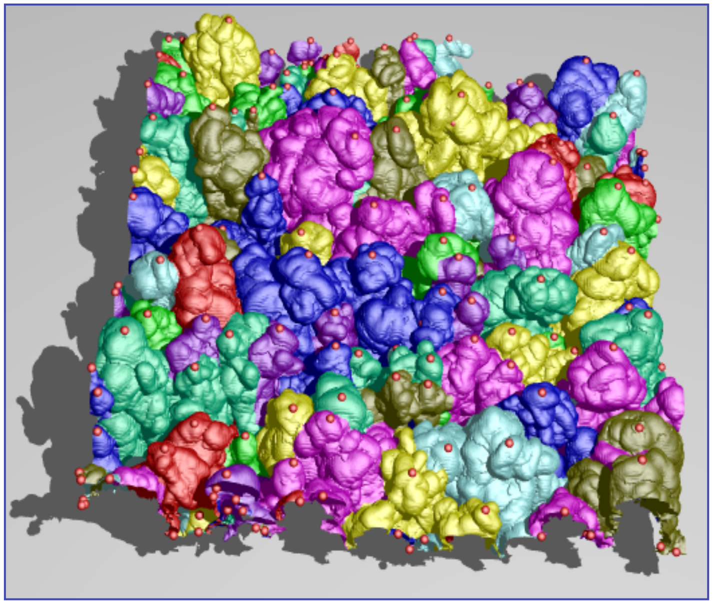
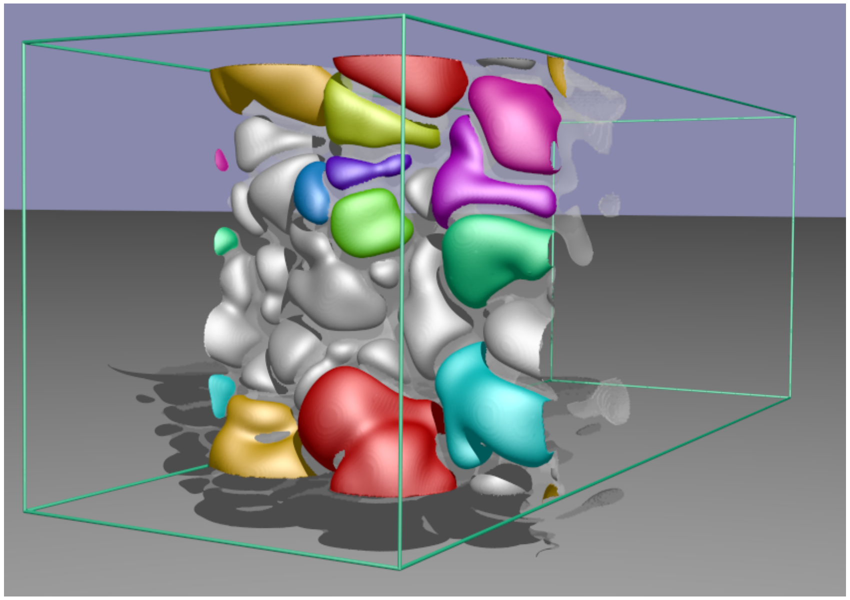
Topological Analysis of Large Scale Data: Theory and Practice



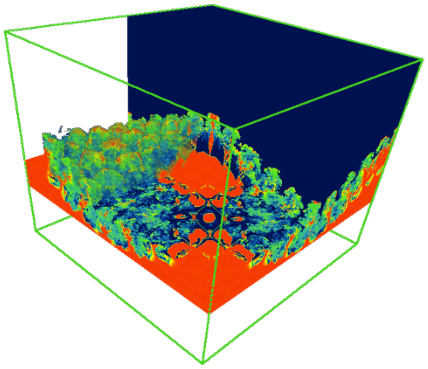


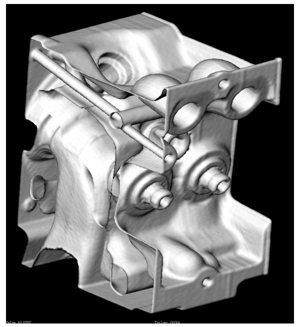
Computation Directorate

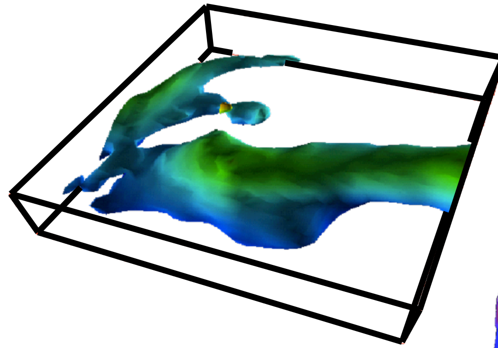
Peer-Timo Bremer

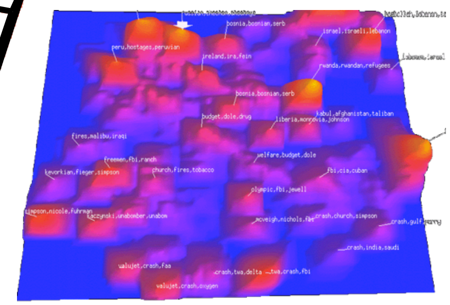
CASC

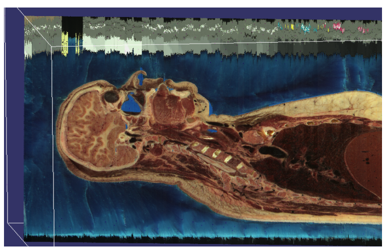
Scientific Data is Only as Useful as the Results From its Analysis



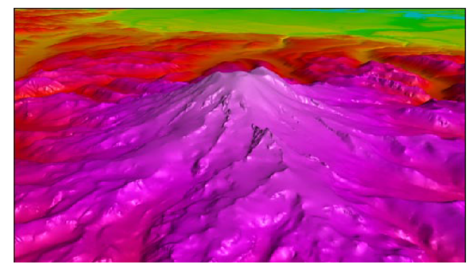


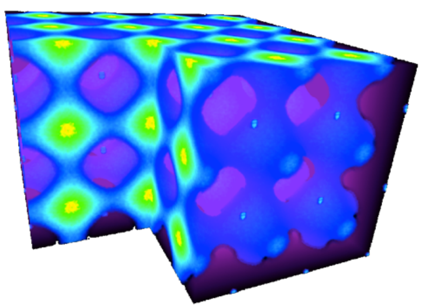


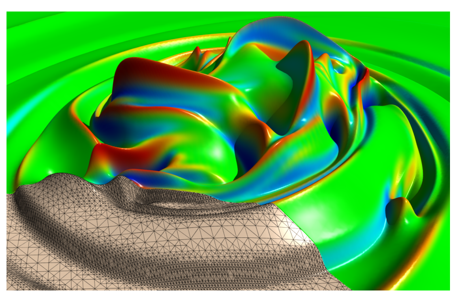


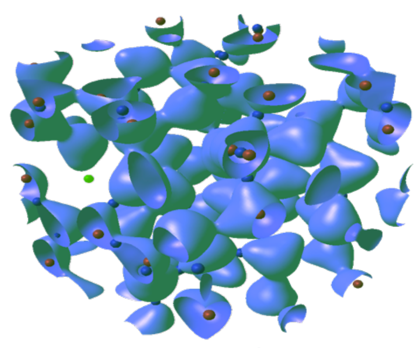


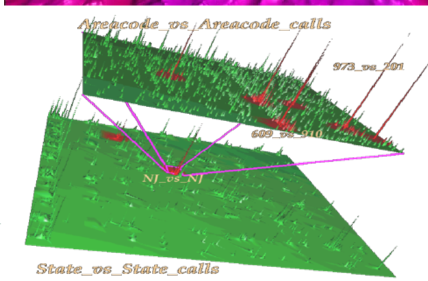
Effectively analyzing scalar functions lies at the Center of a wide variety of application areas









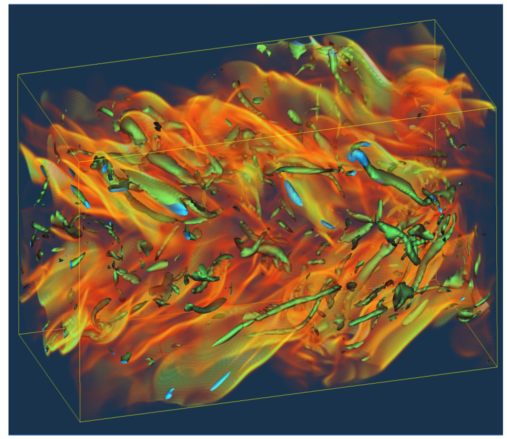


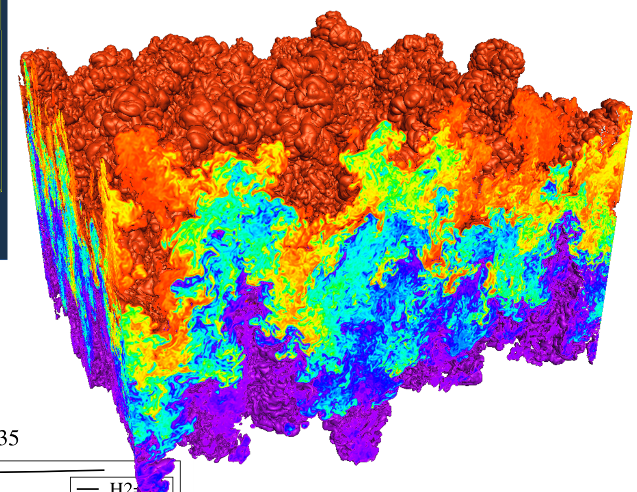
Science has Progressed From Data Poor to Data Rich Leading to Progressively More Complex Analysis Tasks

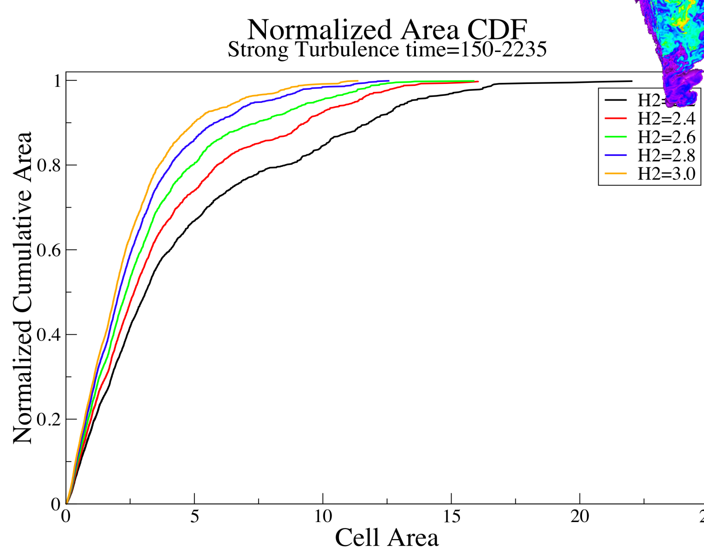
- Hundreds of data points:
  - Tables or simple plots → data collection
- Thousands of data points:
  - Data regression → global trends
- Millions of data points:
  - Data sub-selection → local trends
- Billions of data points:
  - Region of interest → feature based statistics

Feature-Based Statistics Pose Several Challenges

- Feature definition:
  - Intuitive descriptions
- Feature extraction:
  - Efficiency
  - Accuracy
  - Flexibility
- Feature interpretation:
  - Complexity
  - Stability
  - Sensitivity

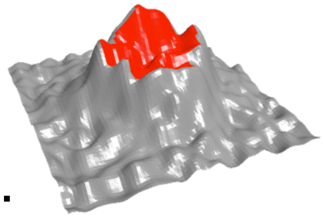


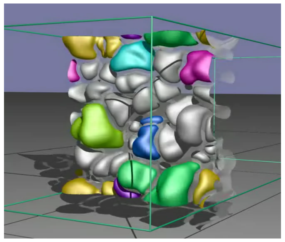




Topology-Based Analysis Techniques Address Some of These Challenges and Enable New Capabilities

- Mathematical theory to define features:
  - Threshold-based features
  - Gradient-based features

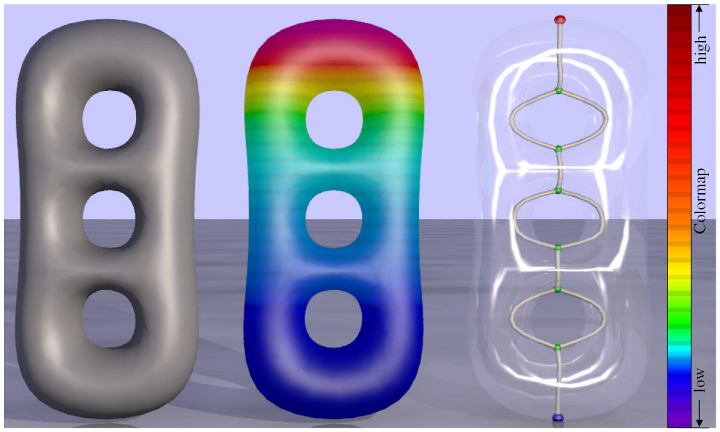


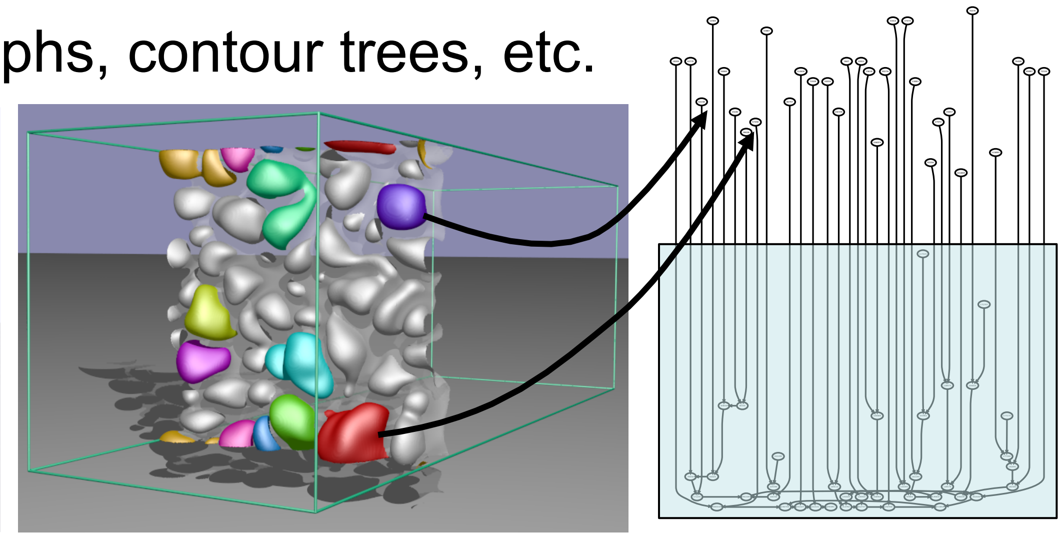


- Efficient, combinatorial algorithms:
  - Generic collection of streaming and/or parallel algorithms
  - Provably correct algorithms and guaranteed error bounds
  - Exact representation of mathematical concepts
- Multi-scale analysis framework with flexible post-processing
  - In-build hierarchical structure with variable simplification metrics
  - Single-pass analysis leading to higher level meta-representations
  - Flexible post-processing of meta-representations enables:
    - parameter selection
    - sensitivity analysis

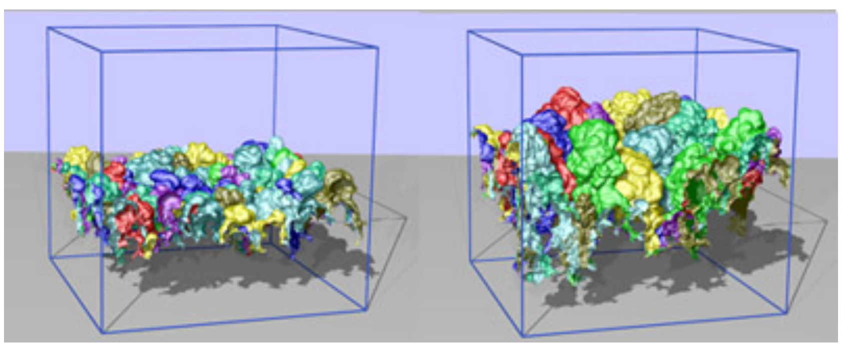
We Have Extended Morse Theory to Analyze the Topology of Discrete Scalar Functions

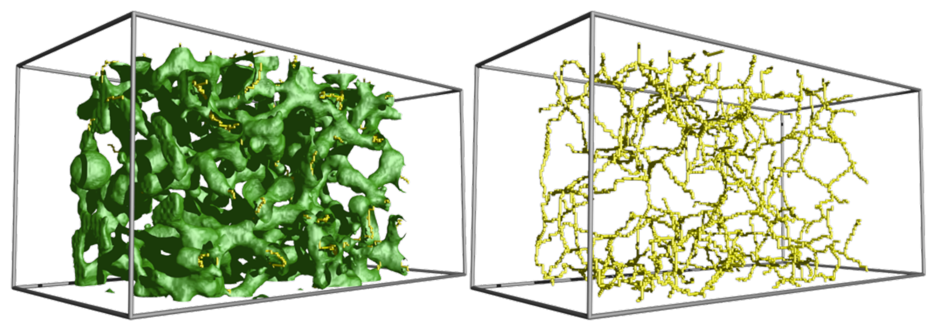
- Level set-based: Reeb graphs, contour trees, etc.





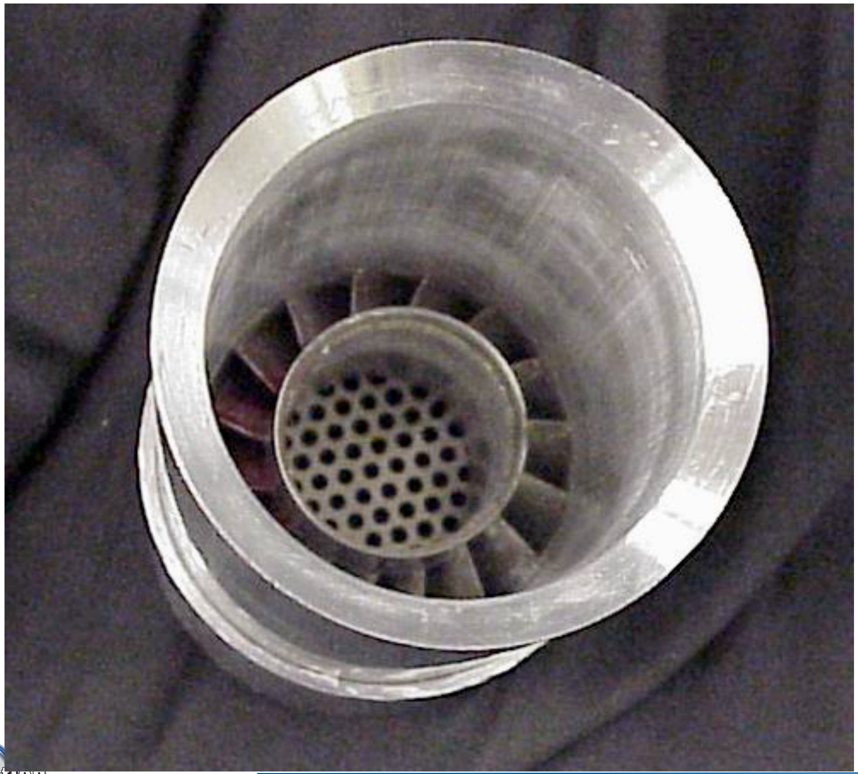
- Gradient-based: Morse-Smale complex





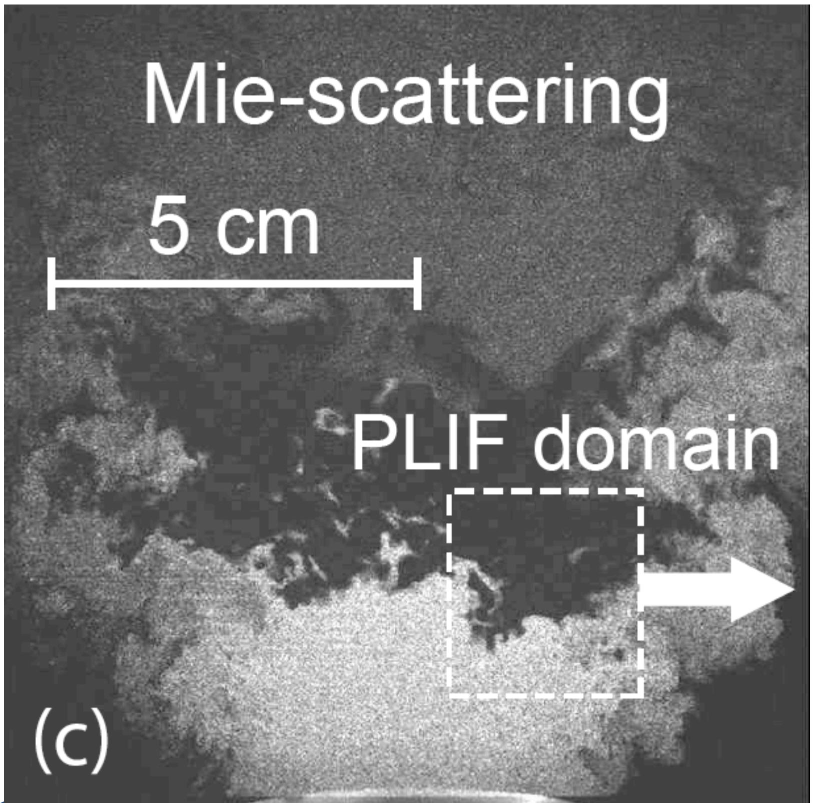
Ultra Clean, Low Swirl Combustion Which Promises Significant Advances in Energy Production Technology

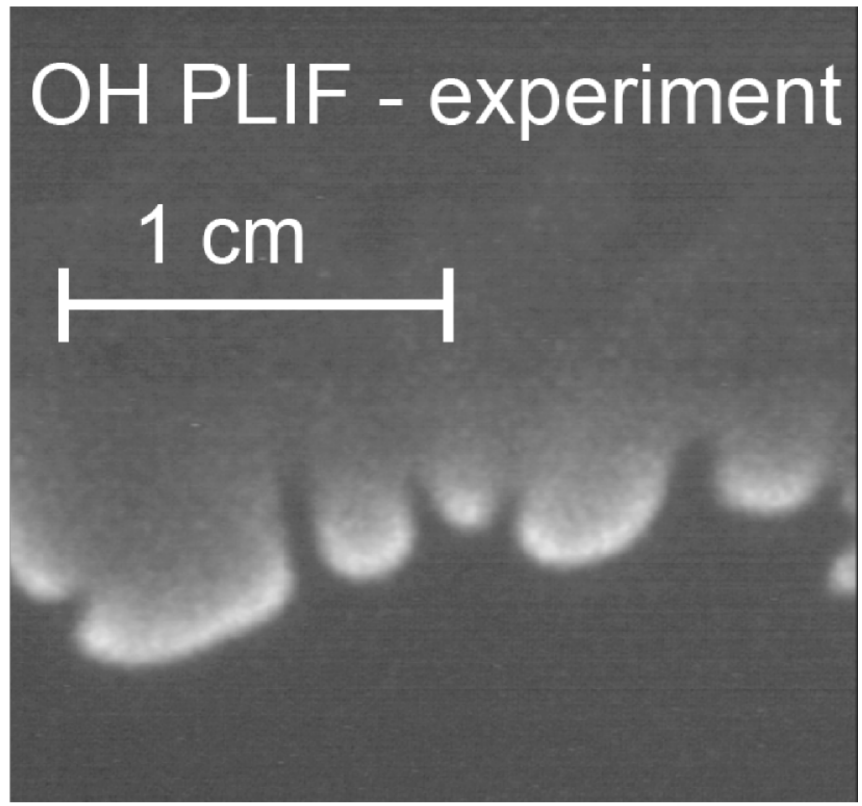
- Low swirl burners produce a stable lifted flame that:
  - Burns more fuel efficient;
  - Produces fewer emissions; and
  - Does not interact with the burner.



Low Swirl, Fuel Lean Flames Burn in a Chaotic, Quasi-Steady Cellular Mode That Defies Traditional Analysis Techniques

- There exists not connected interface separating fuel from the products.
- No notion of a “progress variable” to analyze the dynamics





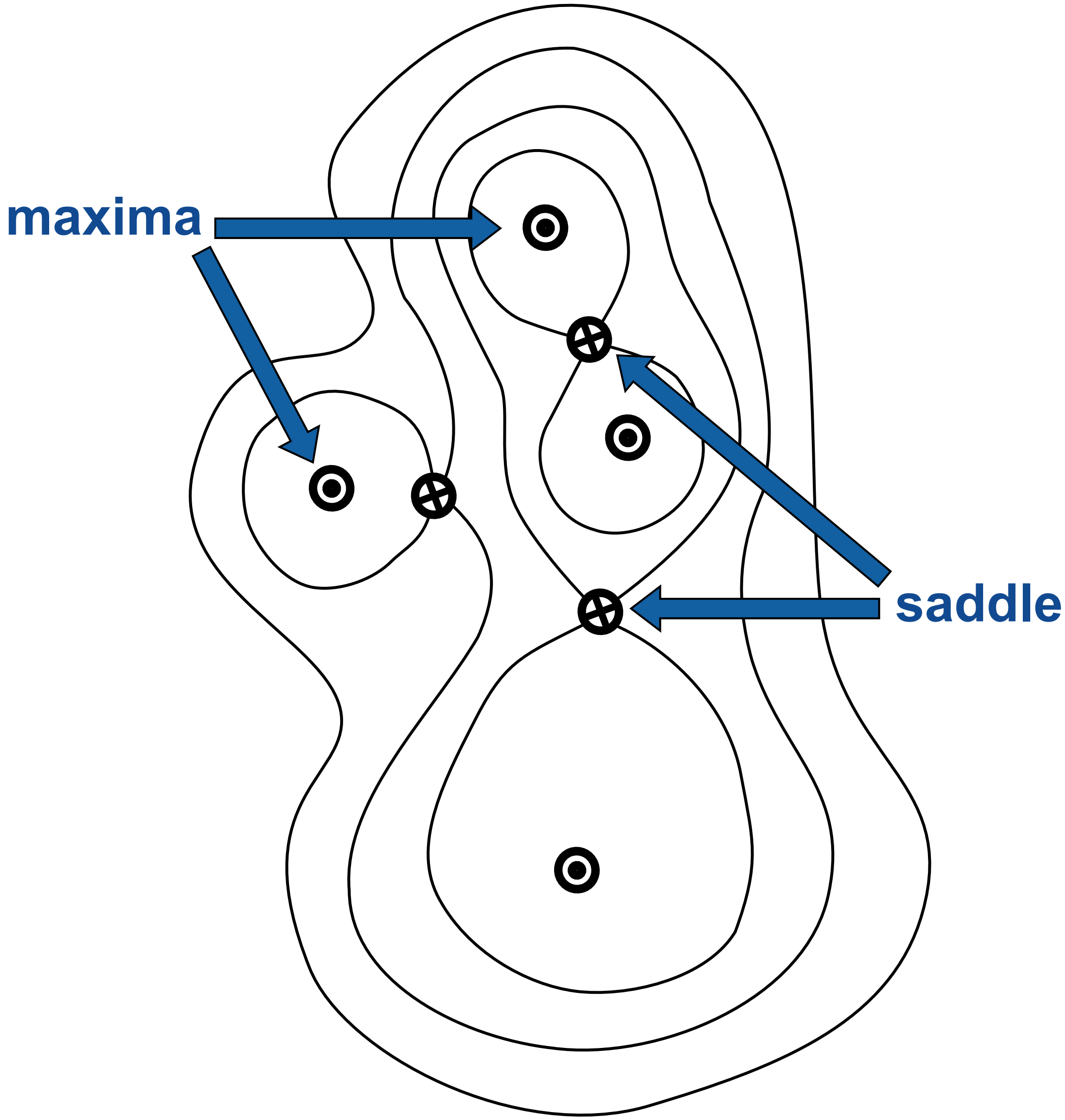
Simulations of Laboratory-Scale Flames are Used to Augment and Validate Experimental Diagnostics

- Simulations allow to:
  - Better interpret diagnostics
  - Form and test new hypothesis
  - Aid in develop new salient models
- We are analyzing two sets of AMR-based simulations of lean, pre-mixed hydrogen flames
  - Three idealized flames at different levels of turbulence
    - 621, 540, and 427 time steps
    - Simulated at effective resolution 512x512x1536
    - 400 GB compressed floating point data
  - Two device scale simulation that differ in flow speed and turbulence
    - 332 and 284 time steps, respectively
    - Simulated at effective resolution 2048^3, saved at 1024^3.
    - 12-16 GB per time step, combined 8.4 TB of raw data

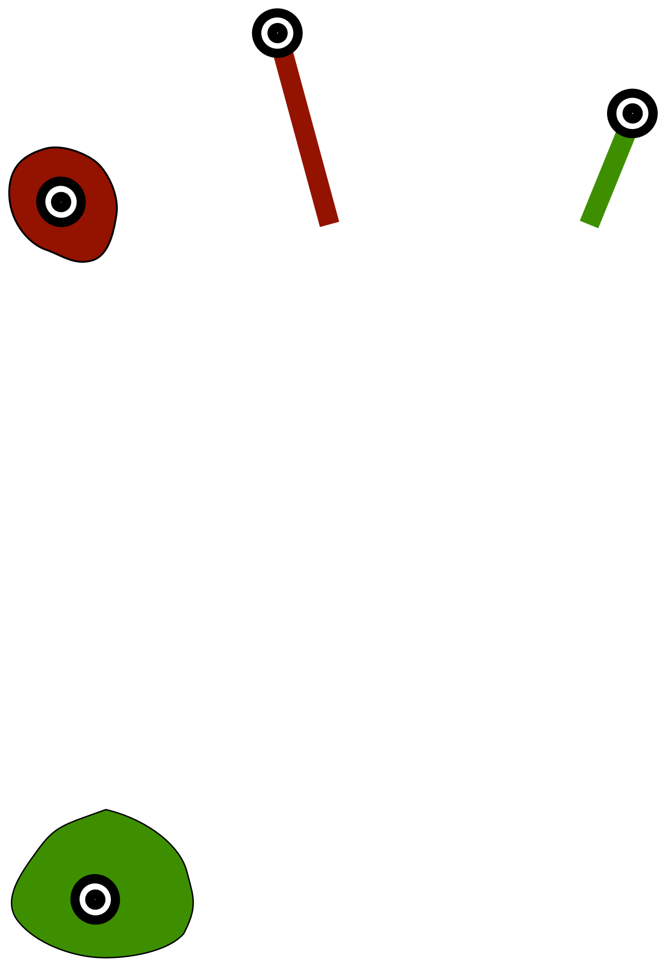
The Features of Primary Interest are Burning Cells Defined via Thresholds of the Local Fuel Consumption

- Scientists characterize the combustion process via burning cells defined as regions of high fuel consumption:
  - How many cells exist at a given time ?
  - What are their sizes ?
  - What are their integral properties, e.g. average temperature ?
  - What is their evolution over time ?
- There exists no pre-set threshold on the fuel consumption and analyzing the data with various thresholds provides important information:
  - How does each of the characteristics change as the threshold changes ?
  - Are there stable thresholds indicating more salient properties ?

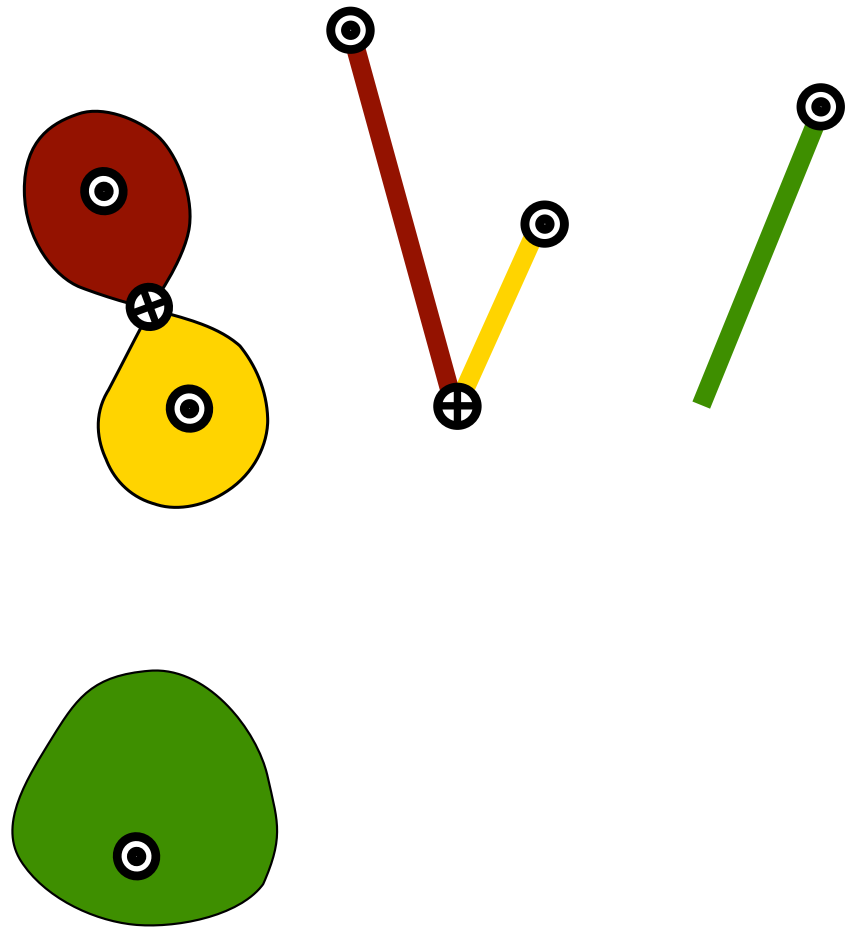
Threshold-Based Features can be Encoded Independent of the Threshold Using Level Set Based Graphs or Trees



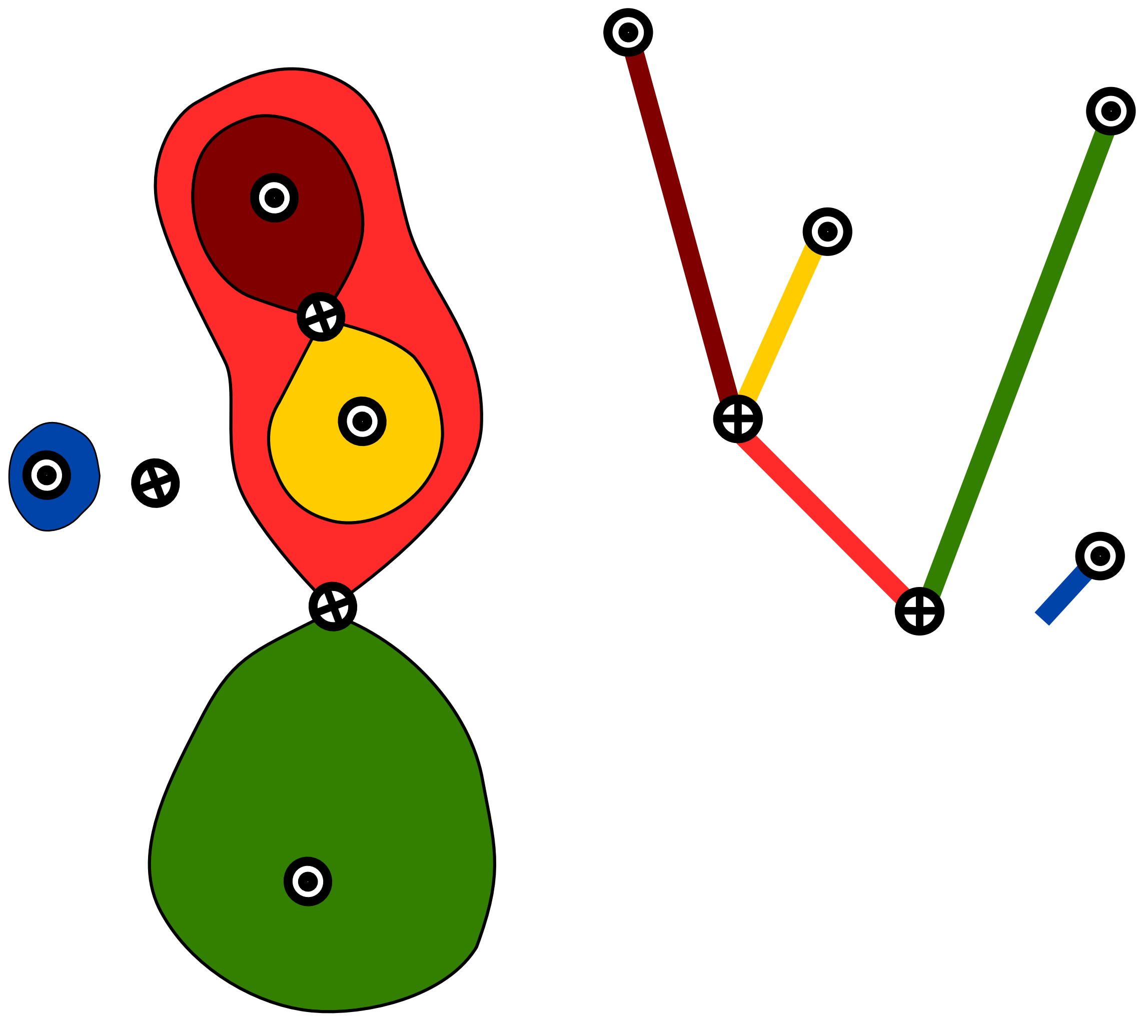
Merge Trees are Ideally Suited to Encode Super-Level Sets: Features Defined as Areas Greater Than a Threshold



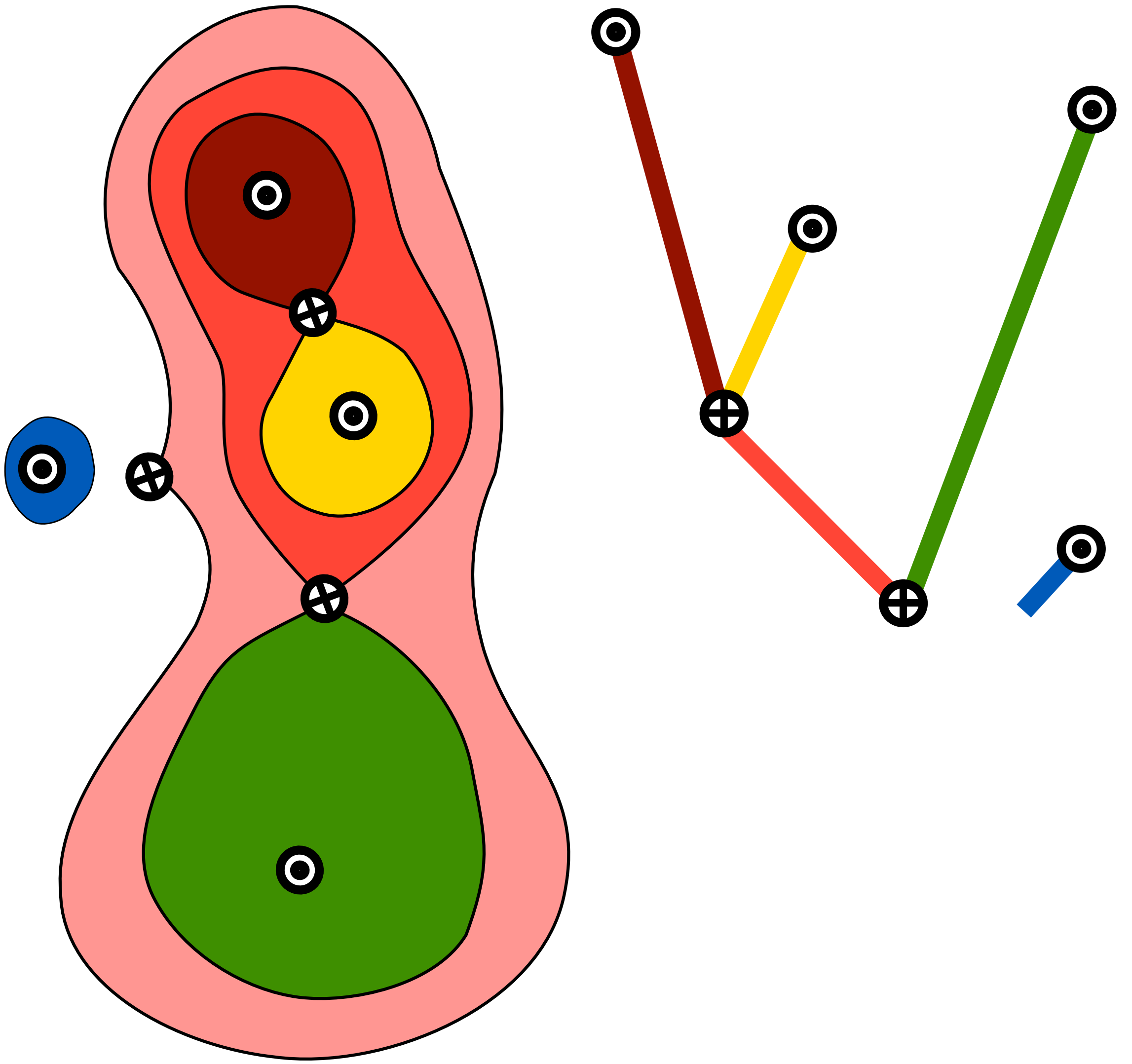
Merge Trees are Ideally Suited to Encode Super-Level Sets: Features Defined as Areas Greater Than a Threshold



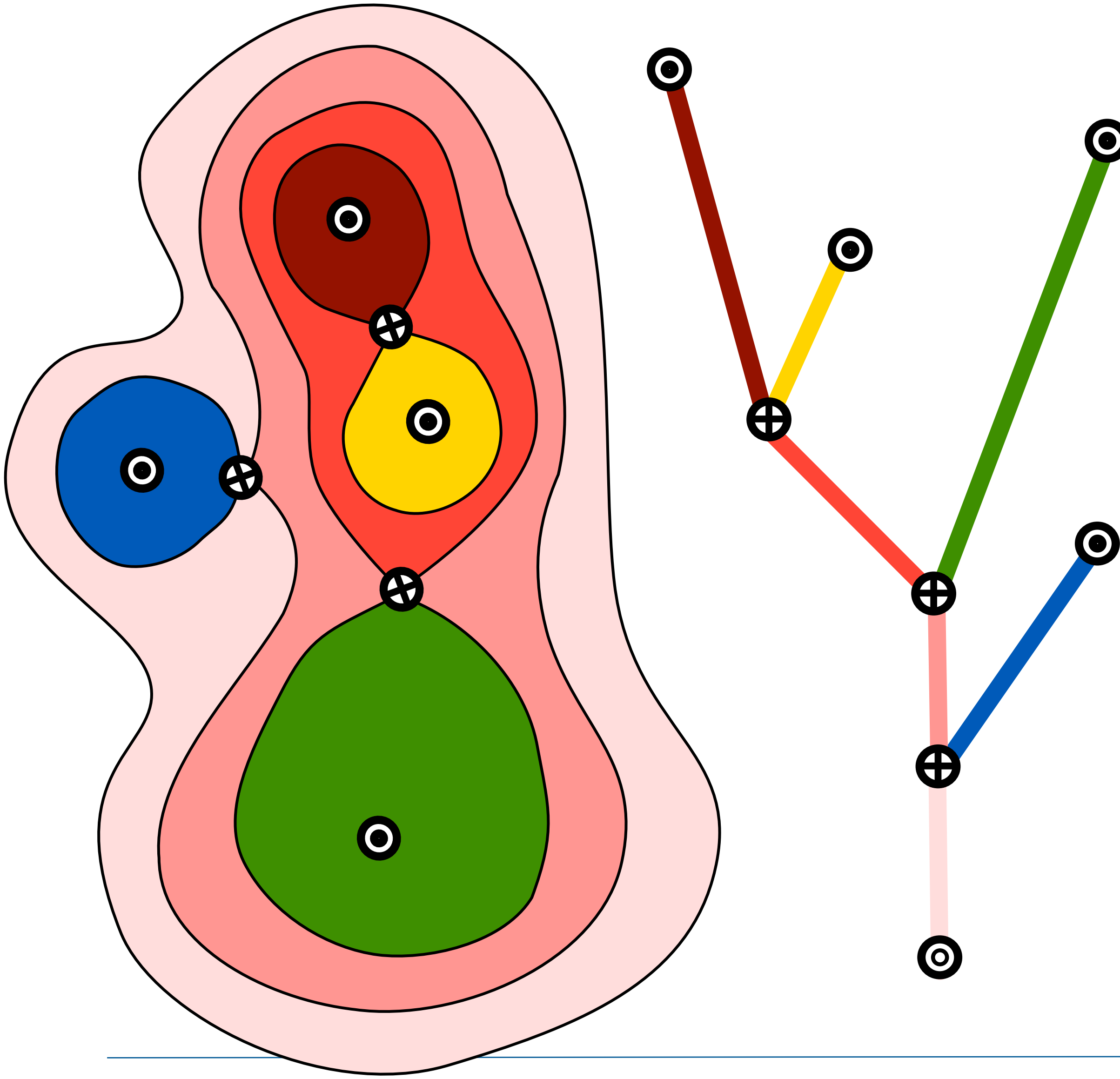
Merge Trees are Ideally Suited to Encode Super-Level Sets: Features Defined as Areas Greater Than a Threshold



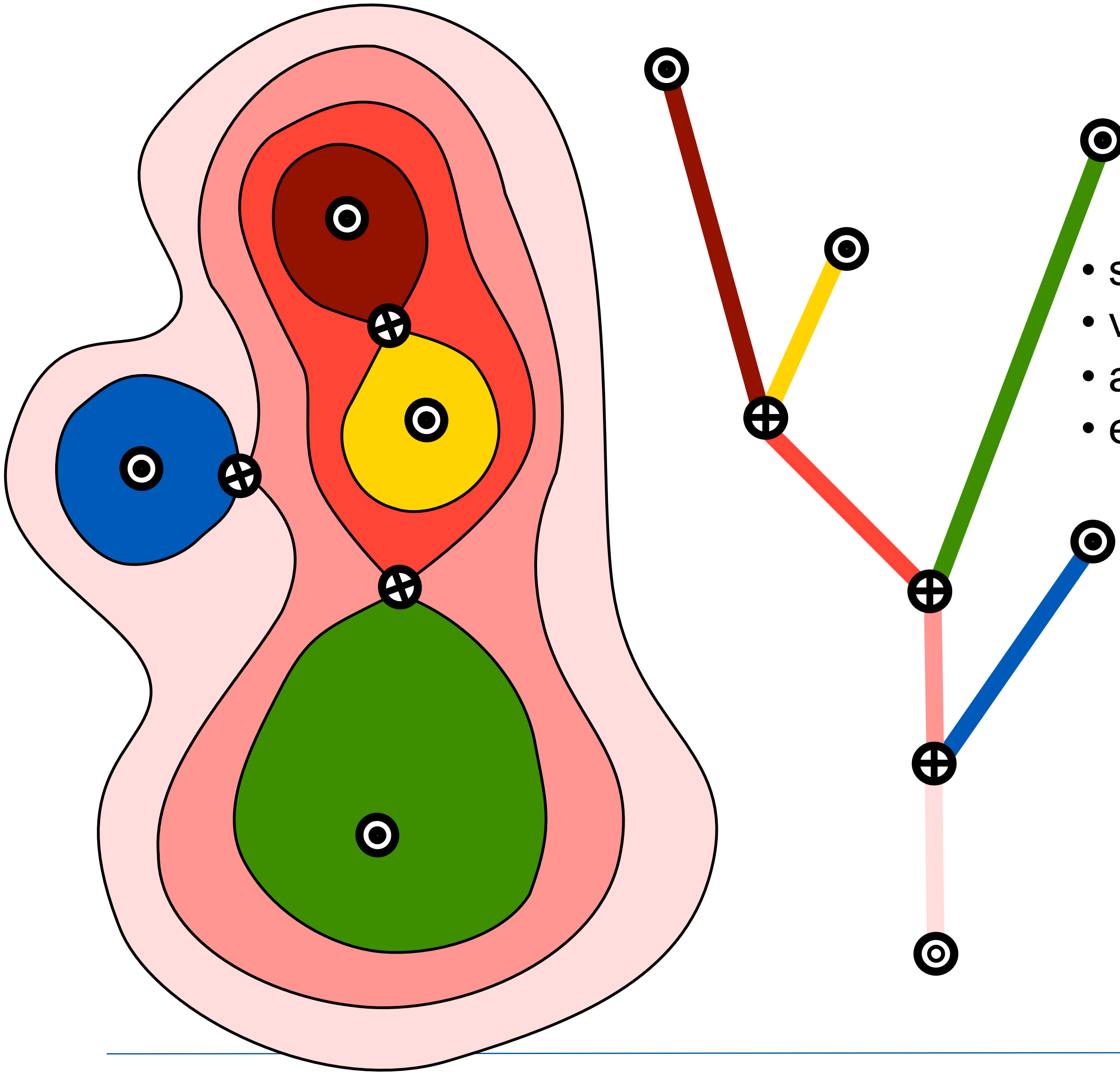
Merge Trees are Ideally Suited to Encode Super-Level Sets: Features Defined as Areas Greater Than a Threshold



Merge Trees are Ideally Suited to Encode Super-Level Sets: Features Defined as Areas Greater Than a Threshold

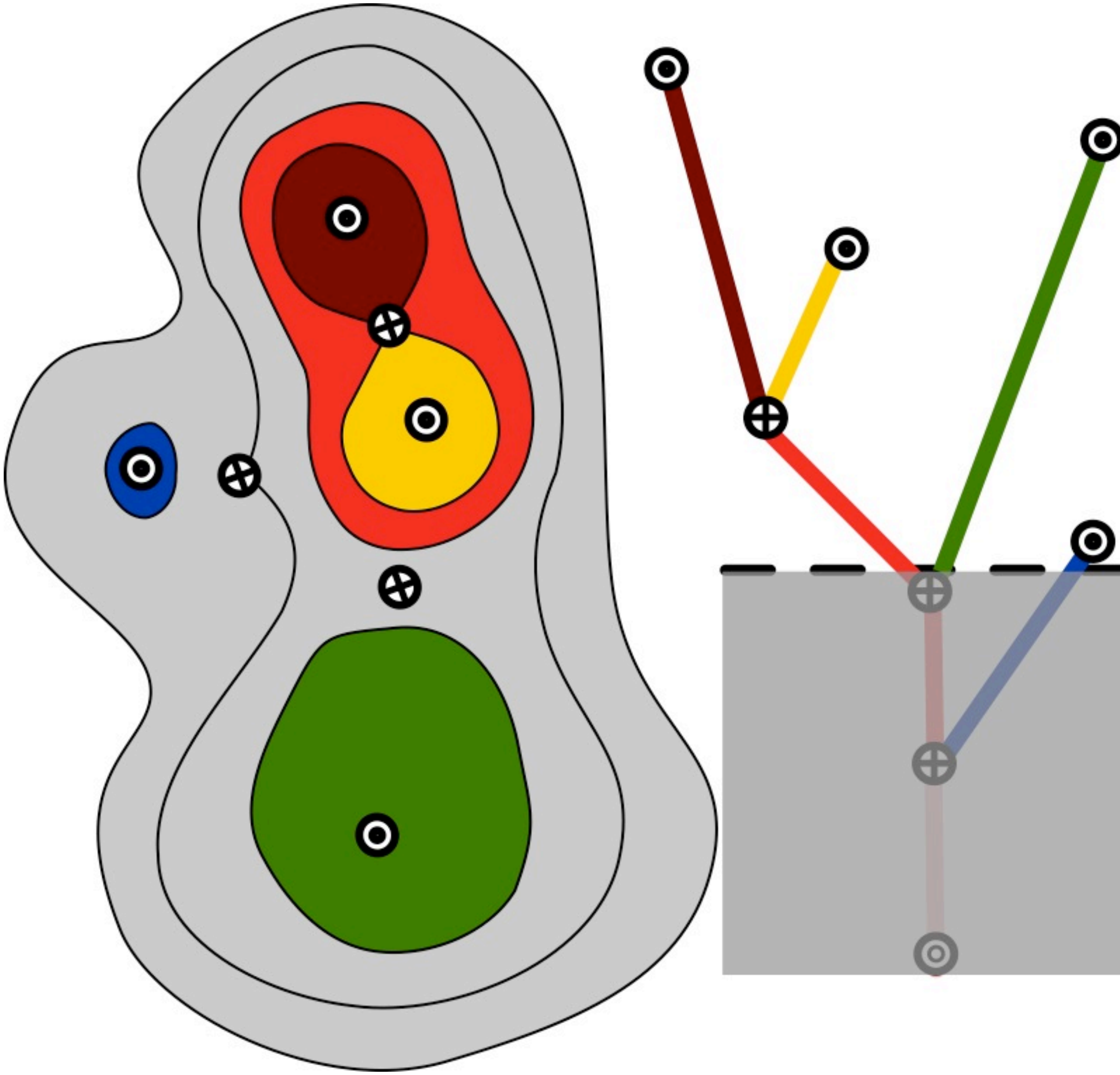


Exploiting the Corresponding Segmentation Merge Trees can Store any Number of Feature Attributes

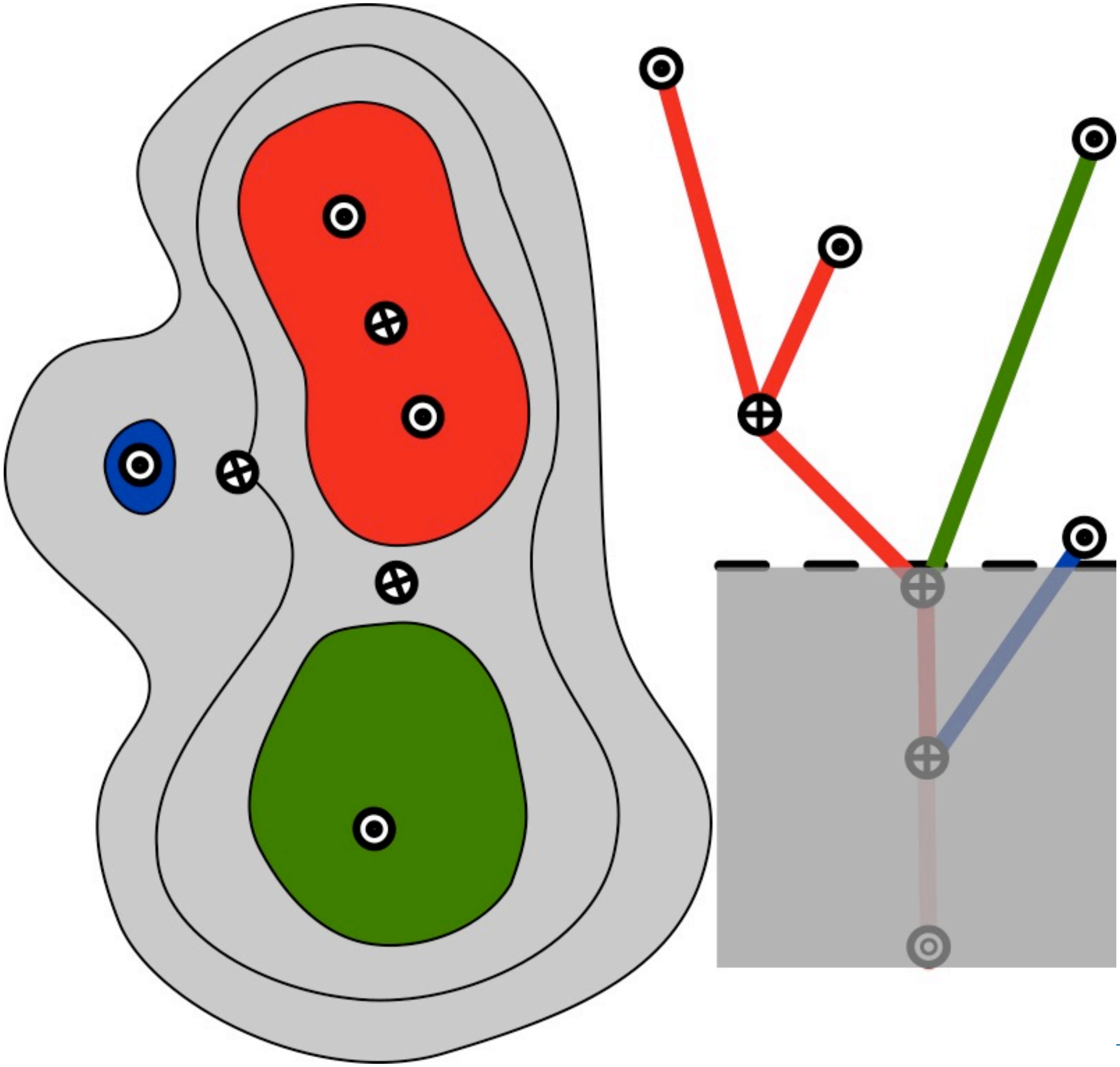


- segmentation
- volume
- average temperature
- etc.

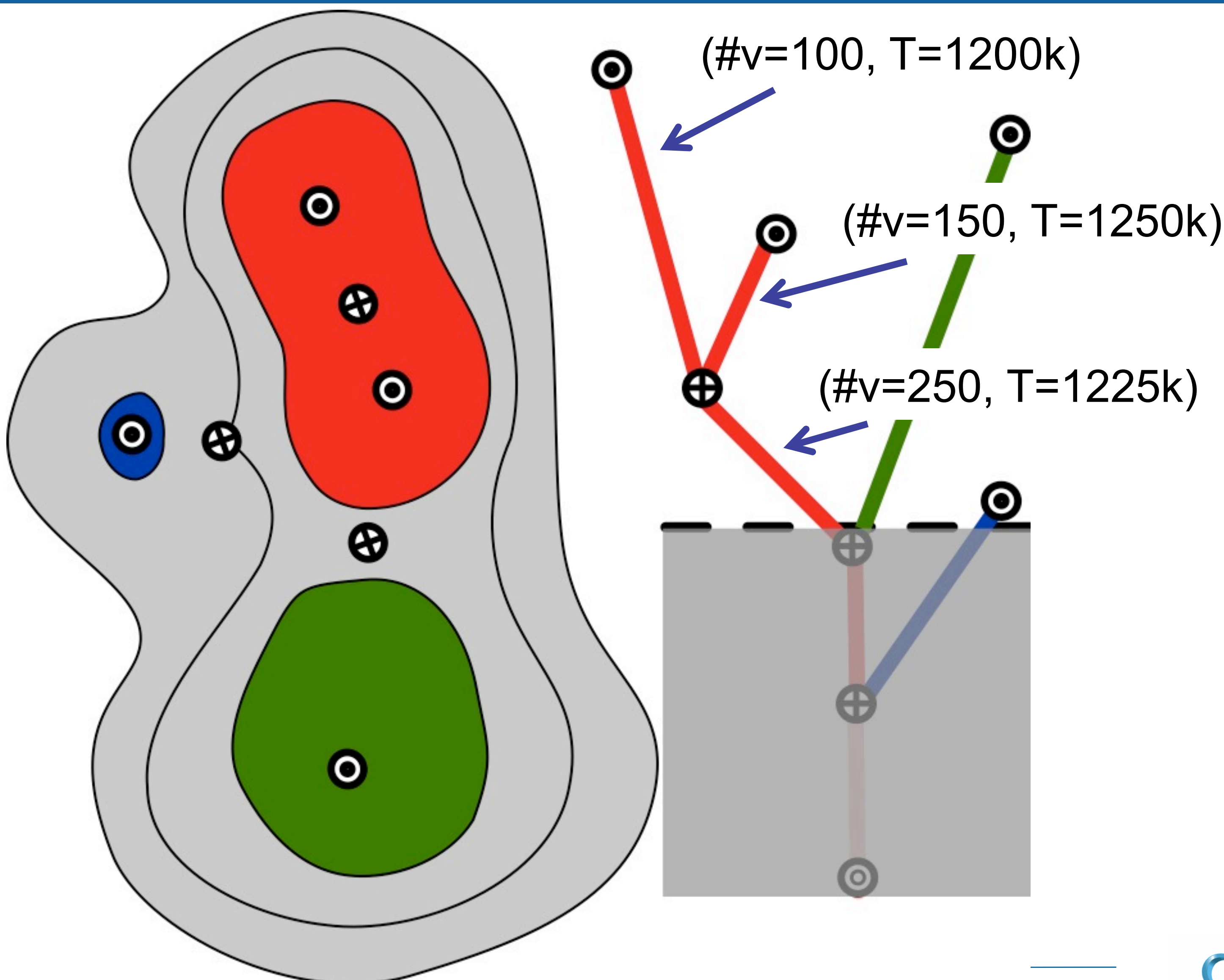
Cutting a Merge Tree Represents Segmentations at Different Thresholds by Treating Combining Sub-Trees



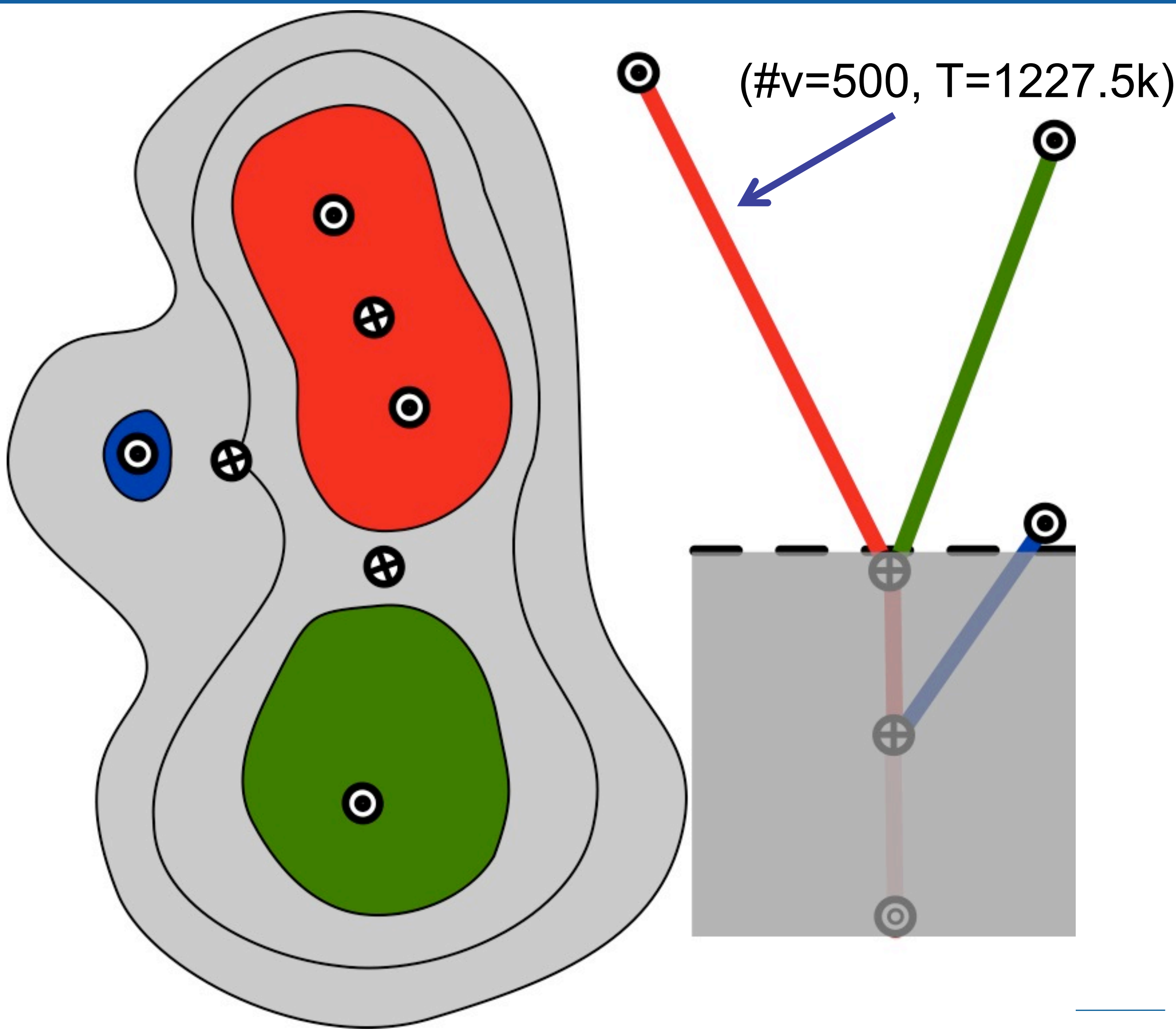
Cutting a Merge Tree Represents Segmentations at Different Thresholds by Treating Combining Sub-Trees



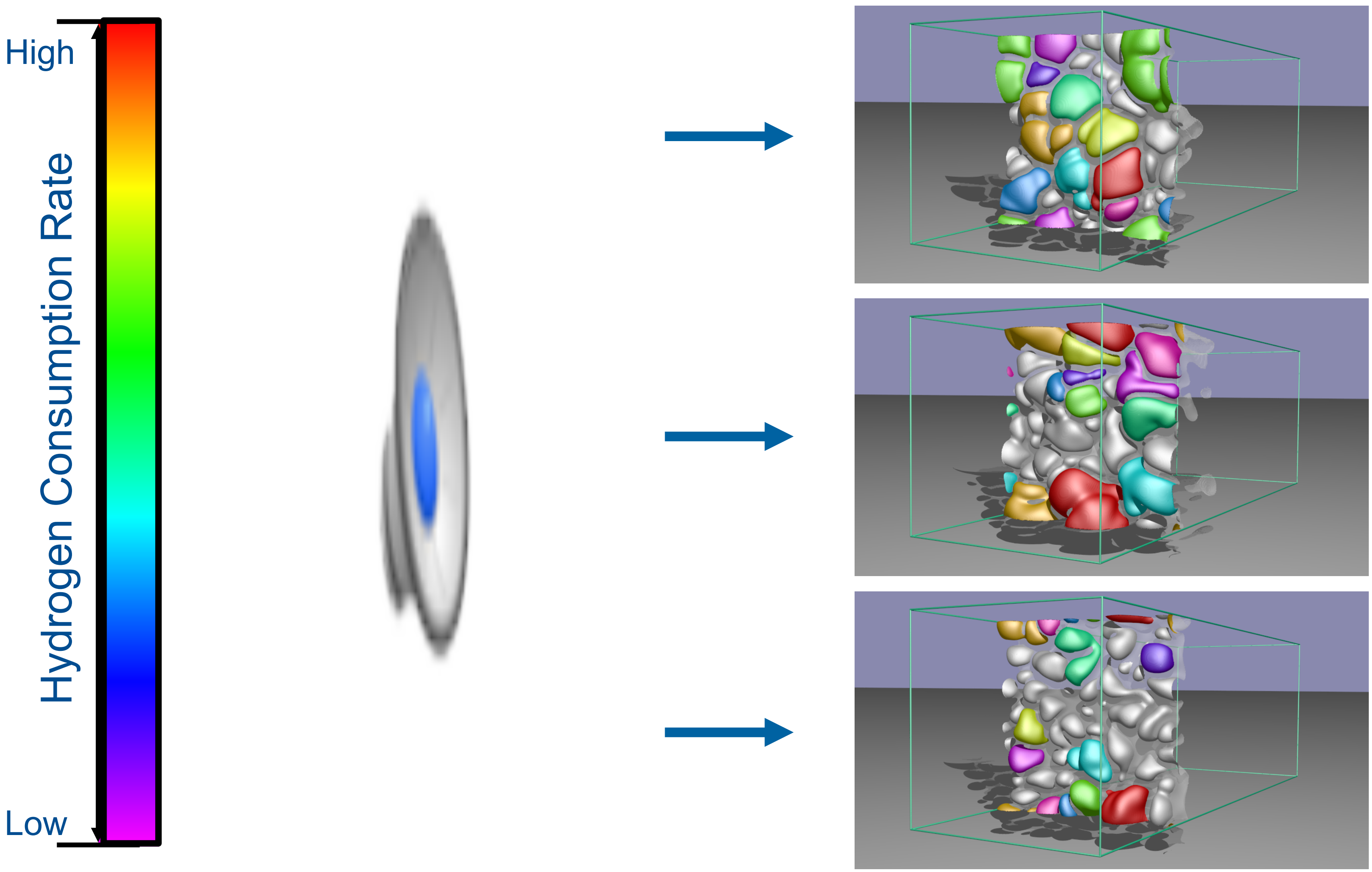
Feature Attributes of the Current Segmentation are Reconstructed Through the Merging of Branches



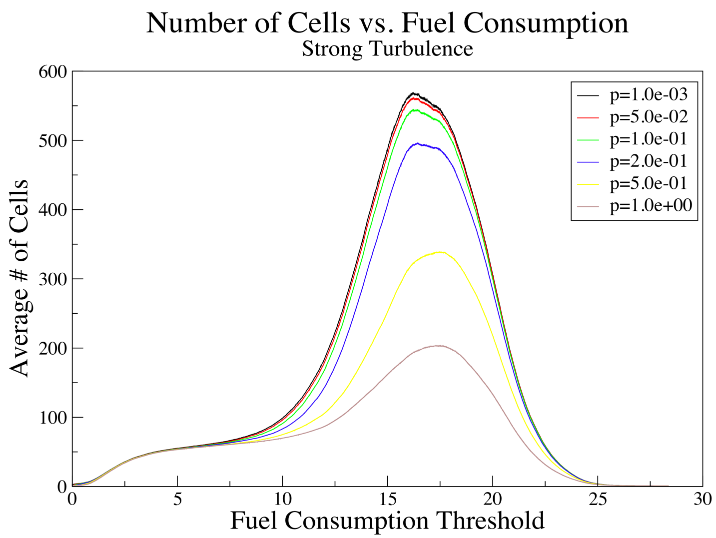
Feature Attributes of the Current Segmentation are Reconstructed Through the Merging of Branches

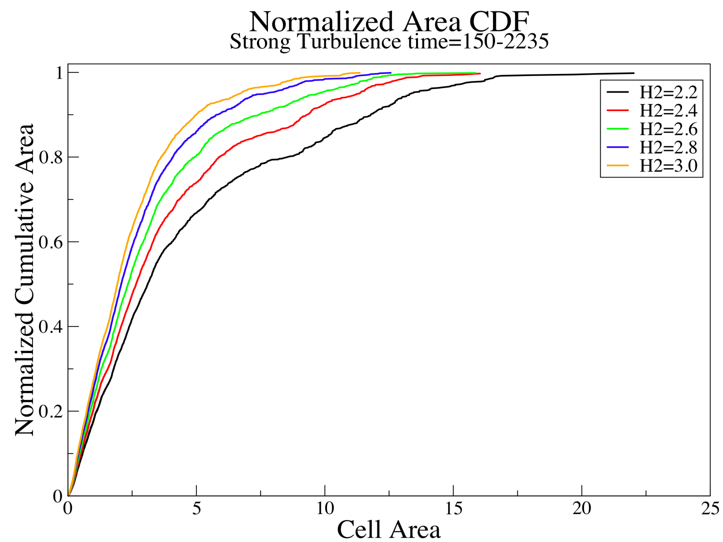


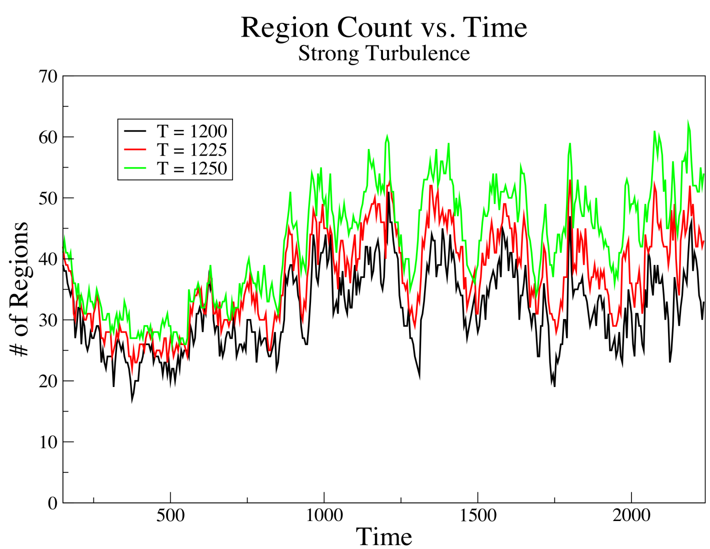
Surface-Based Analysis of Lean, Pre-Mixed Hydrogen in Idealized Conditions Under Different Levels of Turbulence

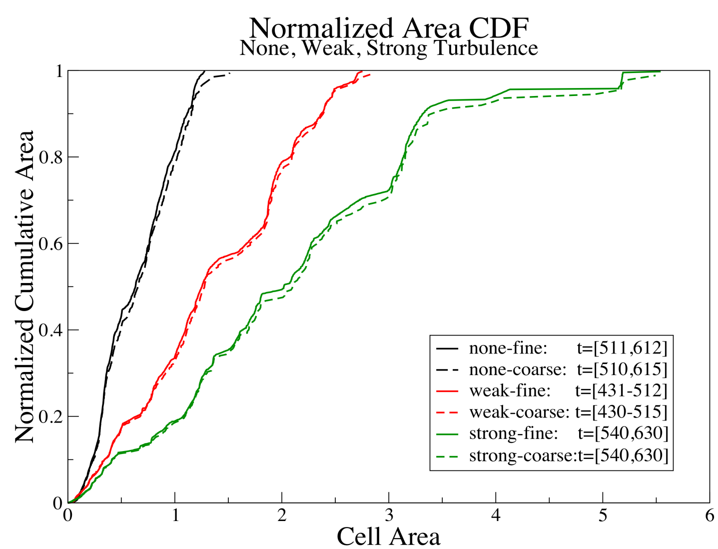


Hierarchical Merge Trees Enable Extensive Parameter Studies in a Single Analysis Path

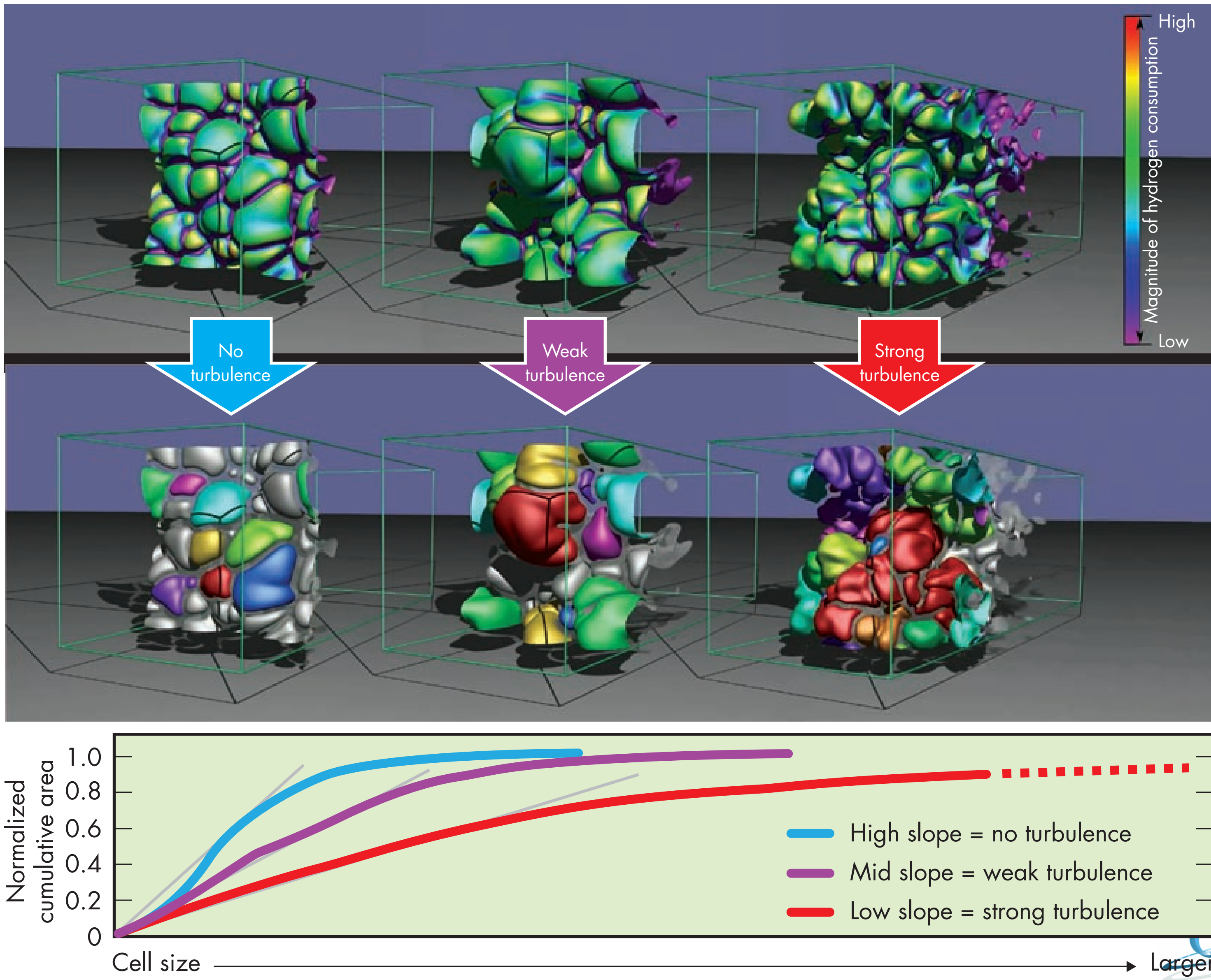




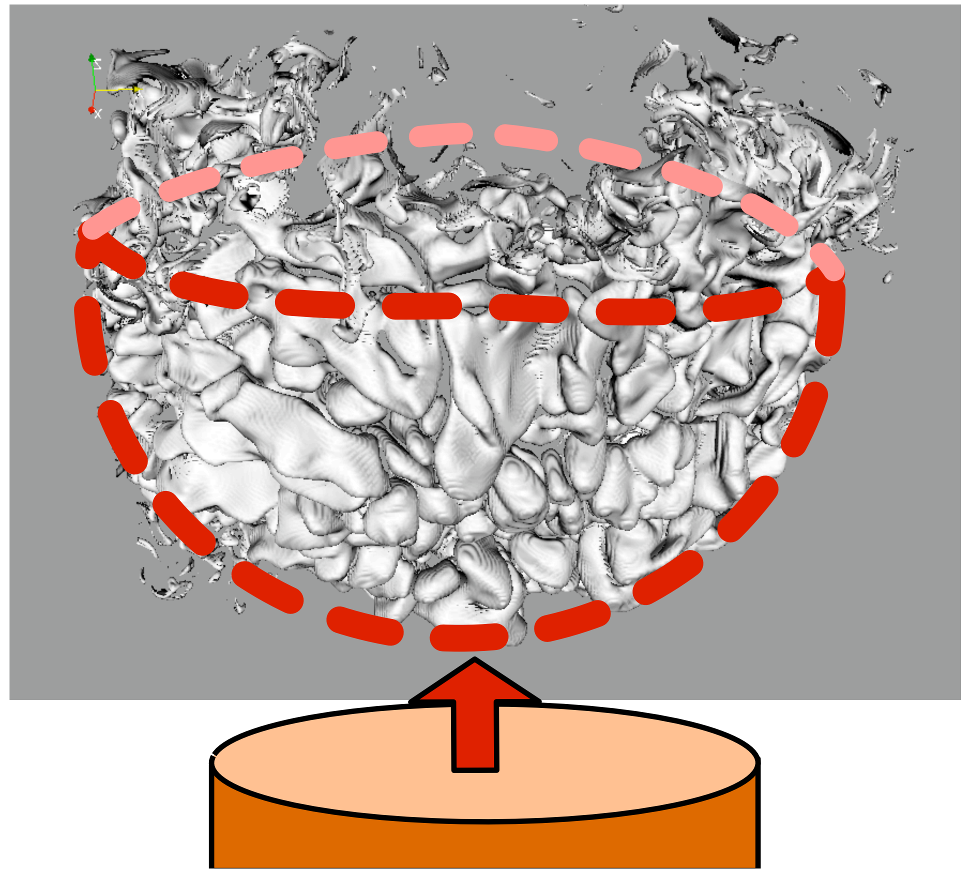




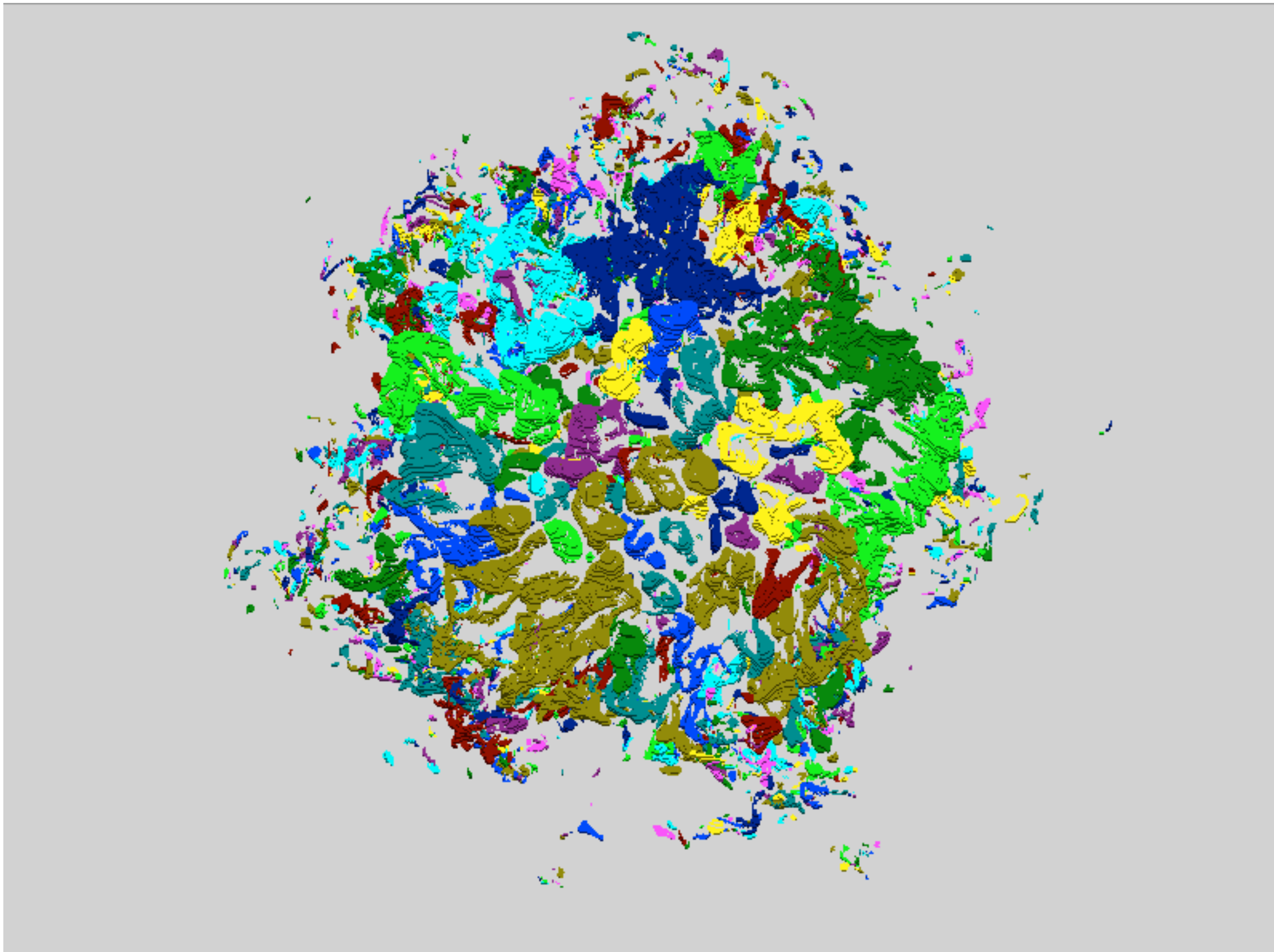
The Analysis Reveals New Scientific Insight Into the Influence of Turbulence on the Combustion Process



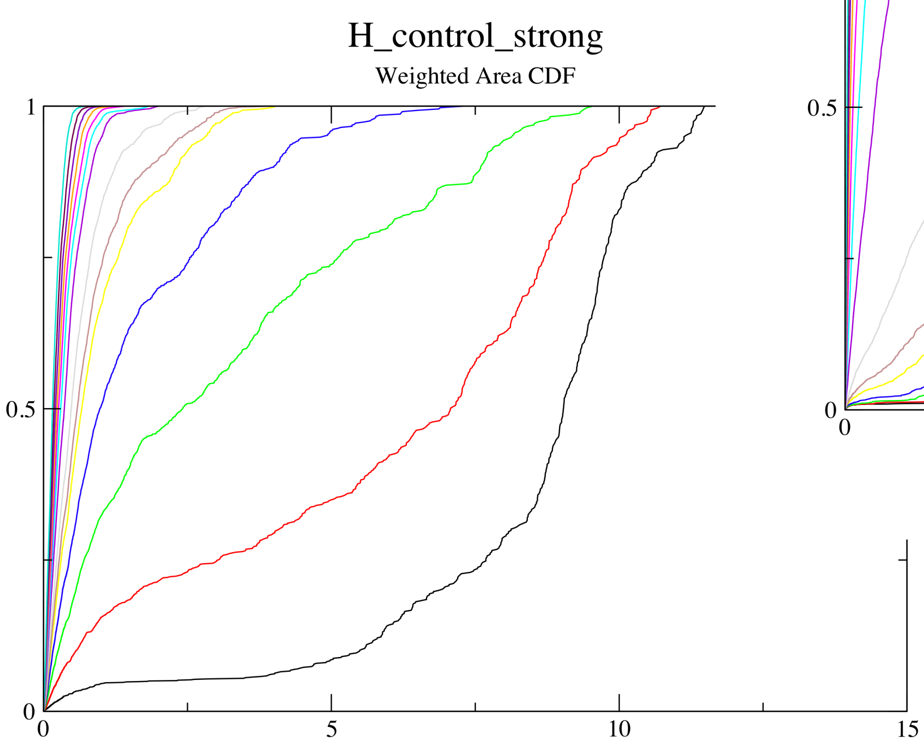
Side-by-Side Comparison on Experimental Methane Flame vs. Burning Regions of a Simulated Hydrogen Flame

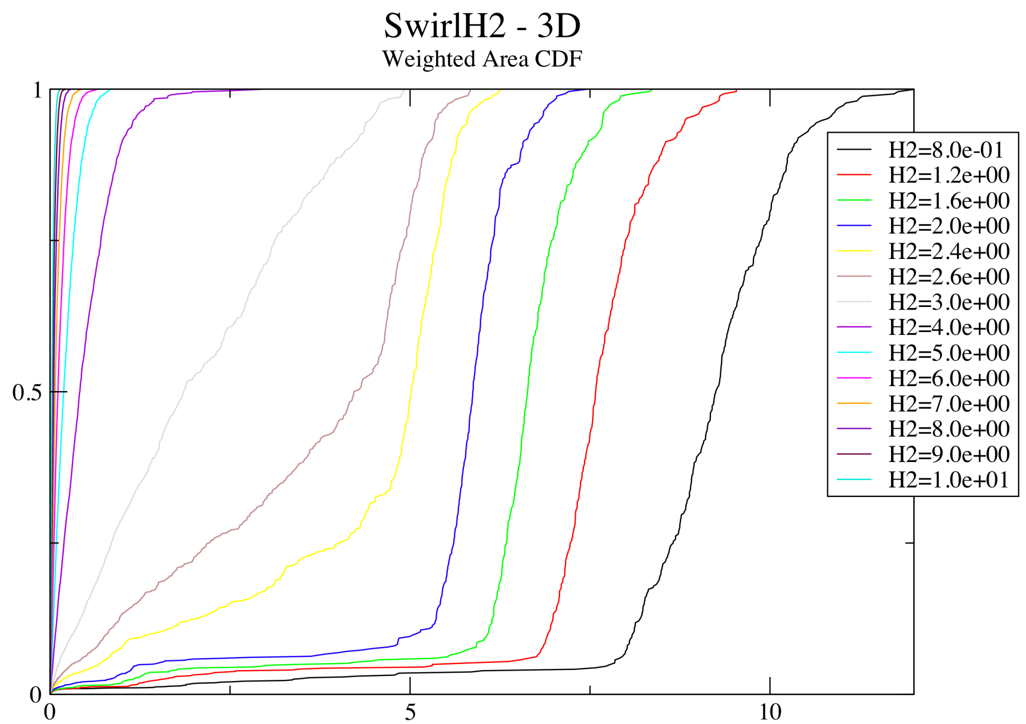


Device-Scale Low-Swirl Turbulent Combustion

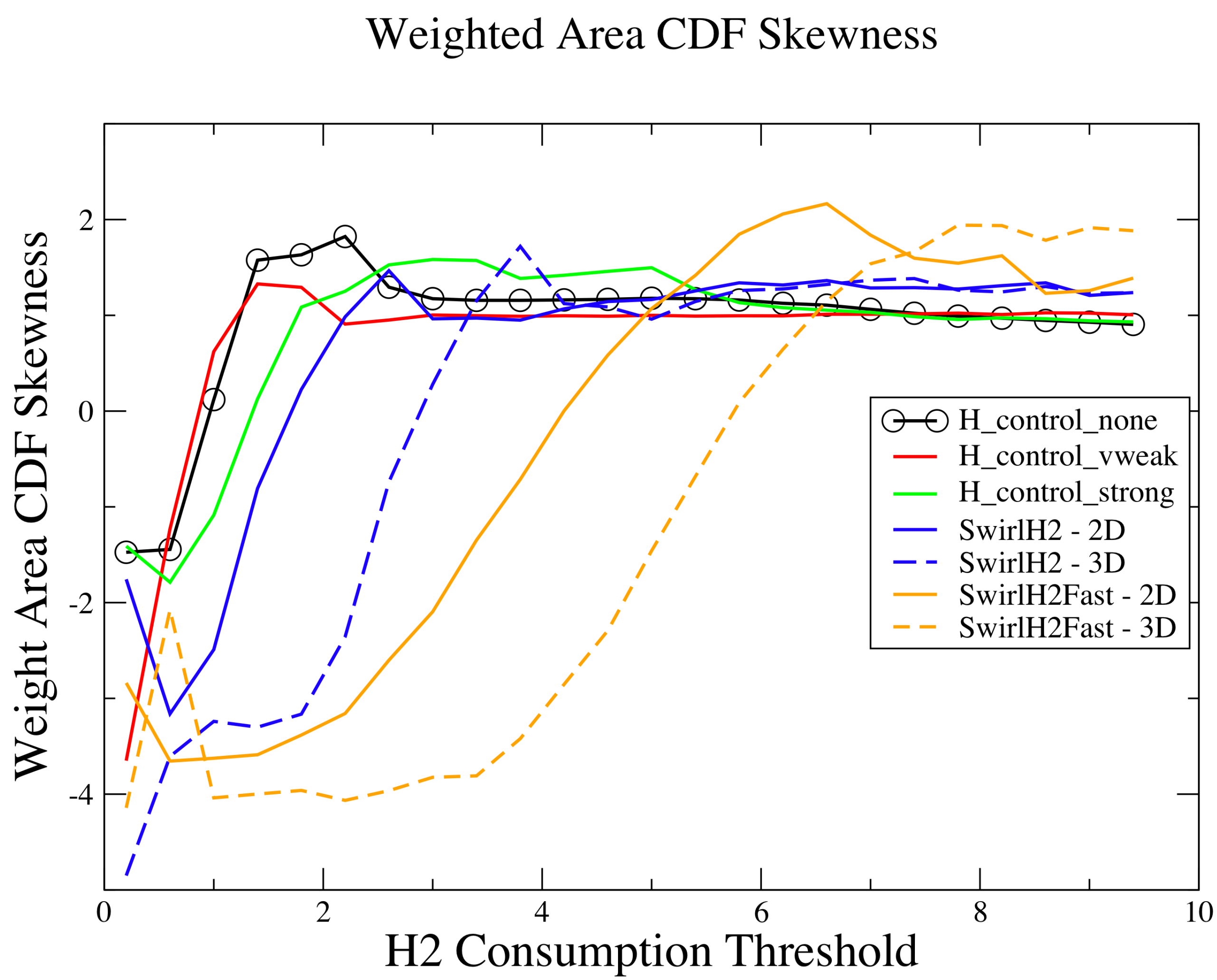


Inter-Simulation Comparison

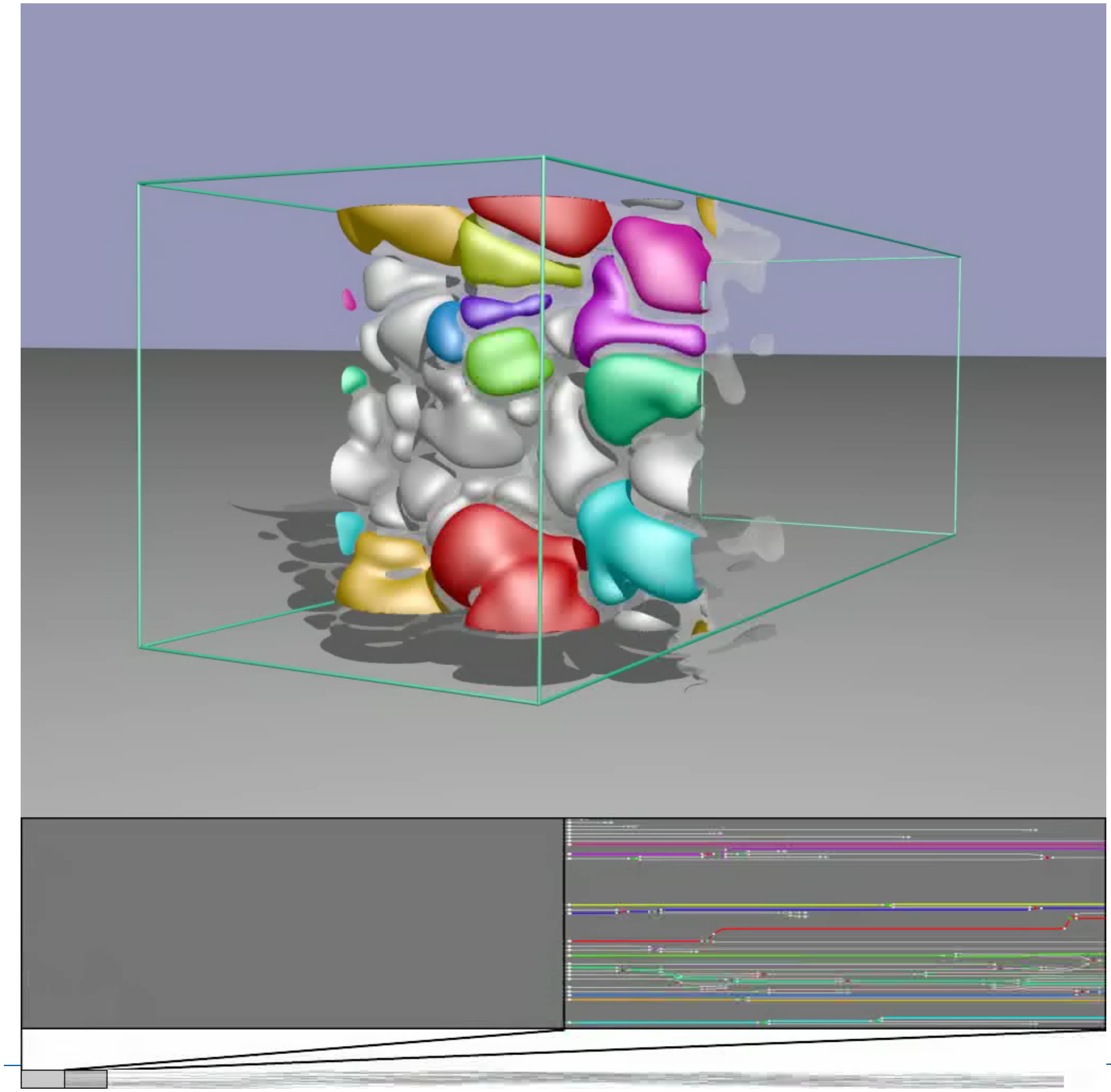




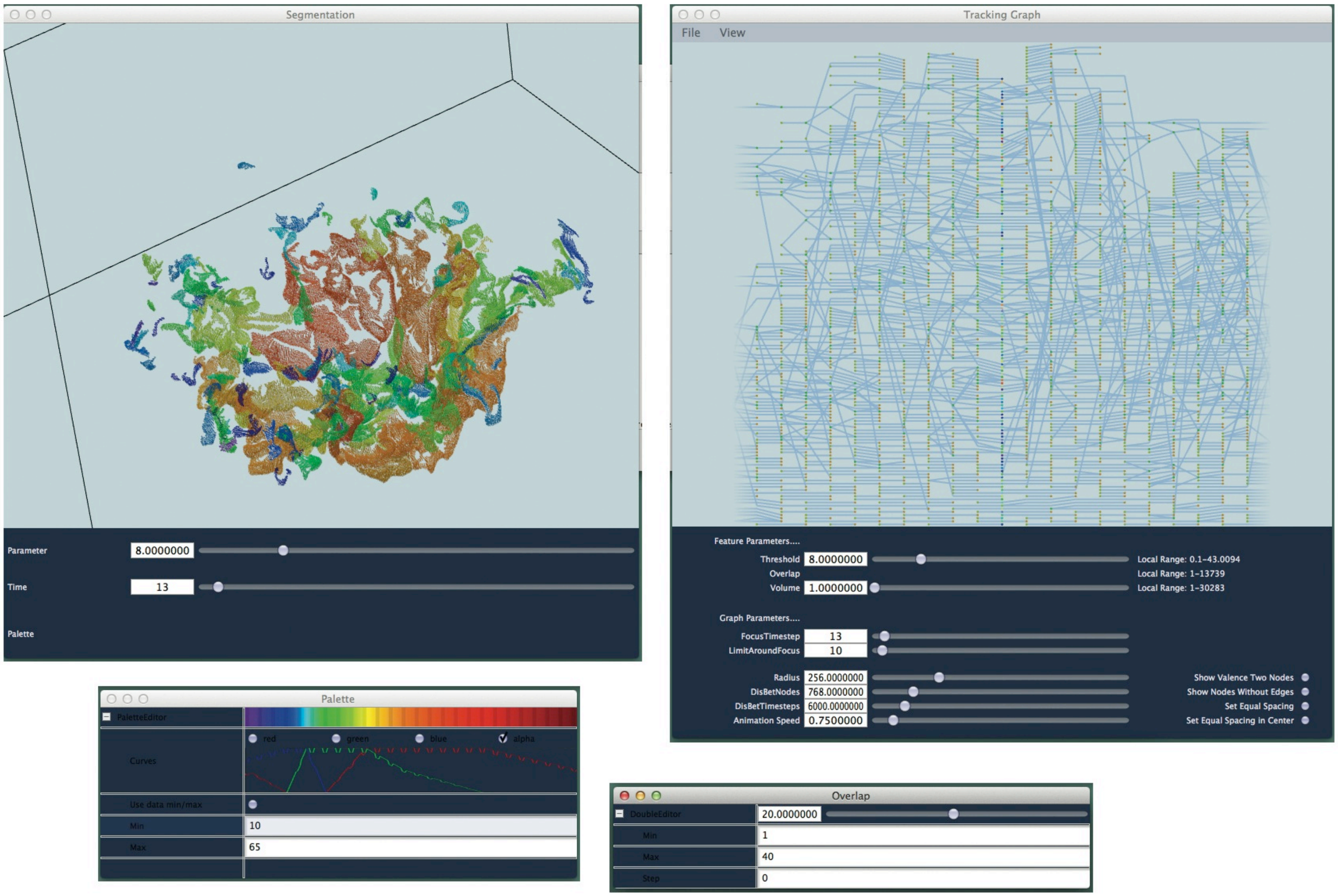
The One-Parameter Family of Segmentations Provides Unprecedented Analysis Capabilities



Overlap-Based Tracking Enables Us to Robustly Track All Cells Through All Time Steps



Recently we have Developed Techniques for On-The-Fly Tracking Producing Dynamic Graph Layouts of Massive Graphs



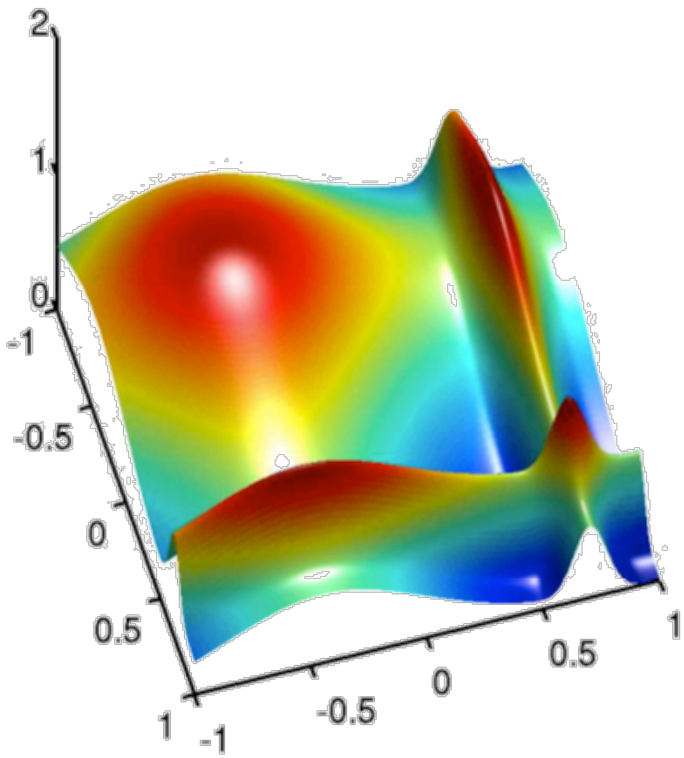
Demo

High-Dimensional Data Represents a Growing Challenge for Visualization and Data Analysis

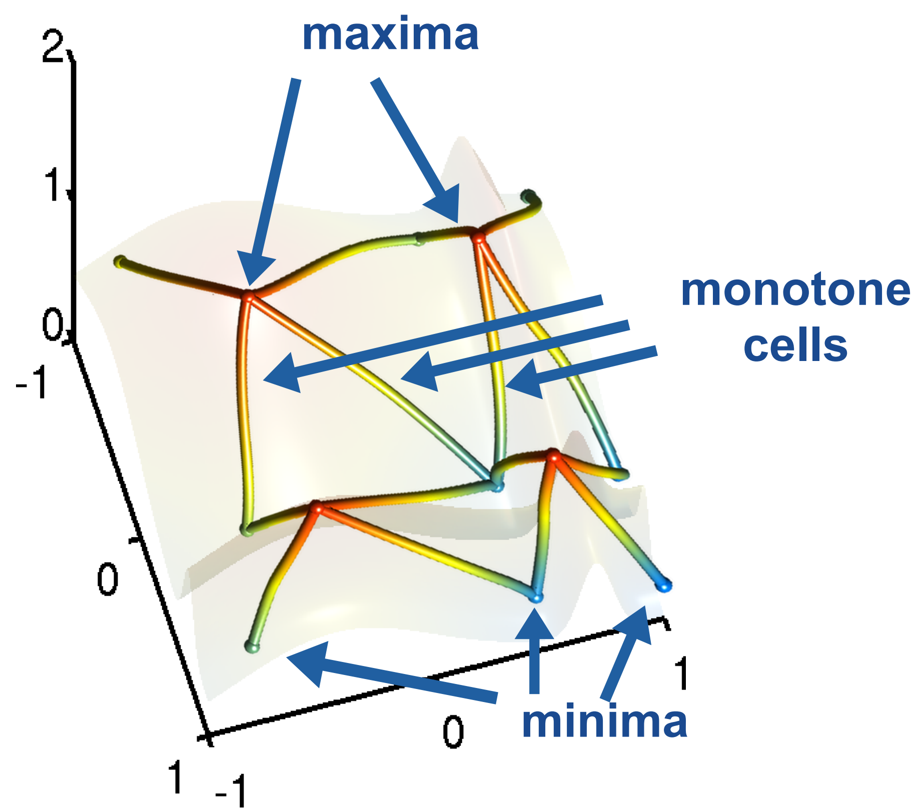
- High dimensional functions represent an increasingly important type of data
  - High dimensional PDF's, i.e. conditional analysis
  - Sensitivity analysis
  - Uncertainty quantification
  - Material phase space
- Current analysis tools are still limited and often unintuitive
  - Scatter plots
  - Parallel coordinates
  - Dimension reduction
  - Clustering
- High-dimensional topological segmentations can provide new opportunities for analysis and new visual metaphors

The High-Dimensional MS Complex Provides a Topological Segmentation Leading to New Analysis and Visualization Techniques

- The Morse-Smale complex encodes the topology of an input function with all extrema and the monotone cells between them



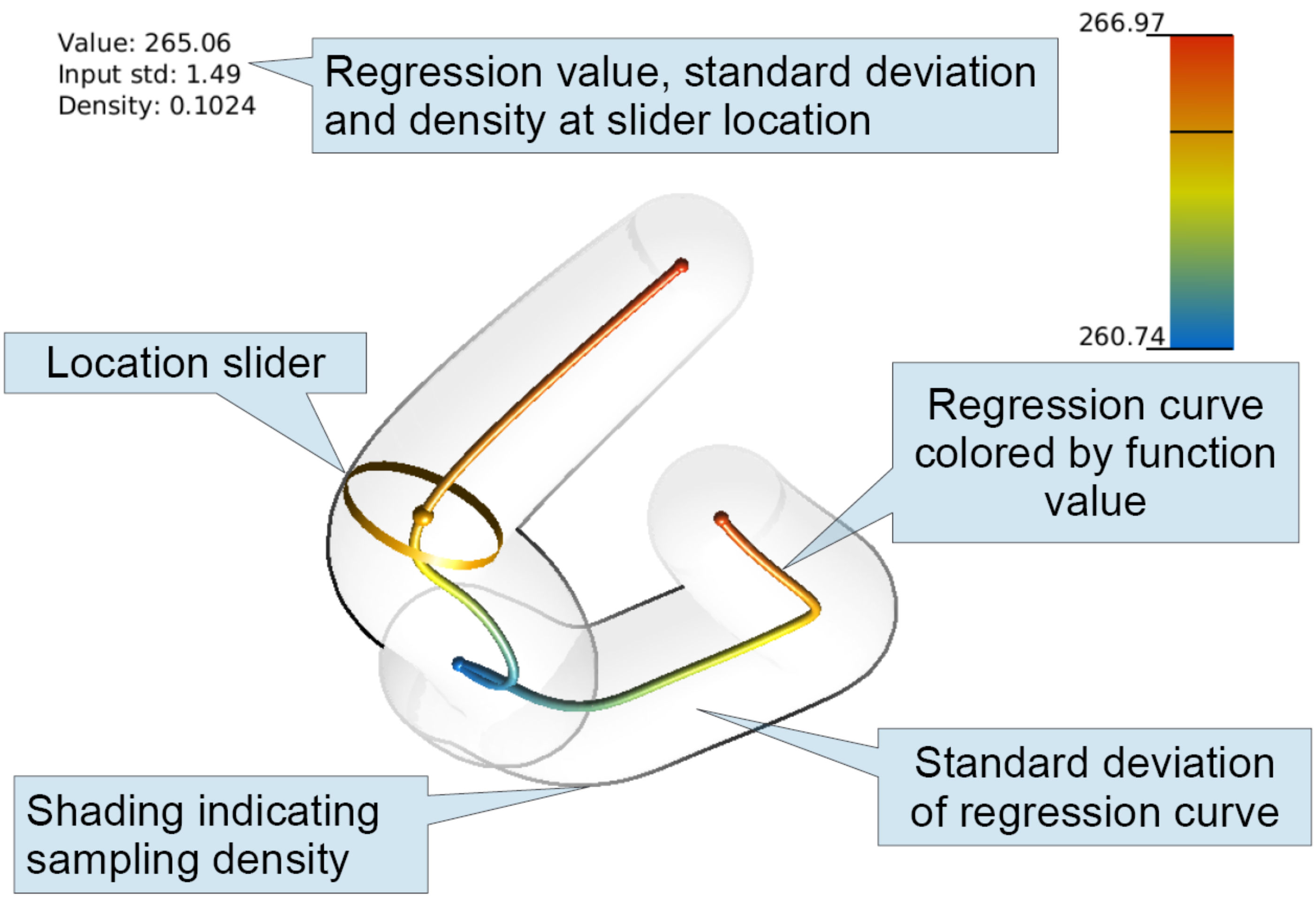
2D test function



dual MS complex from random sample

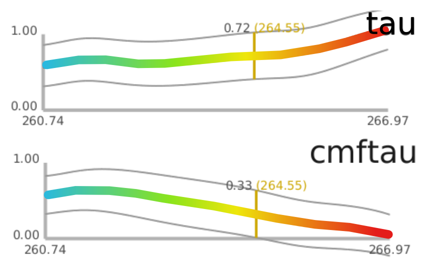
Exploration of High Dimensional Functions for Sensitivity Analysis

Integrated presentation of statistics and topology

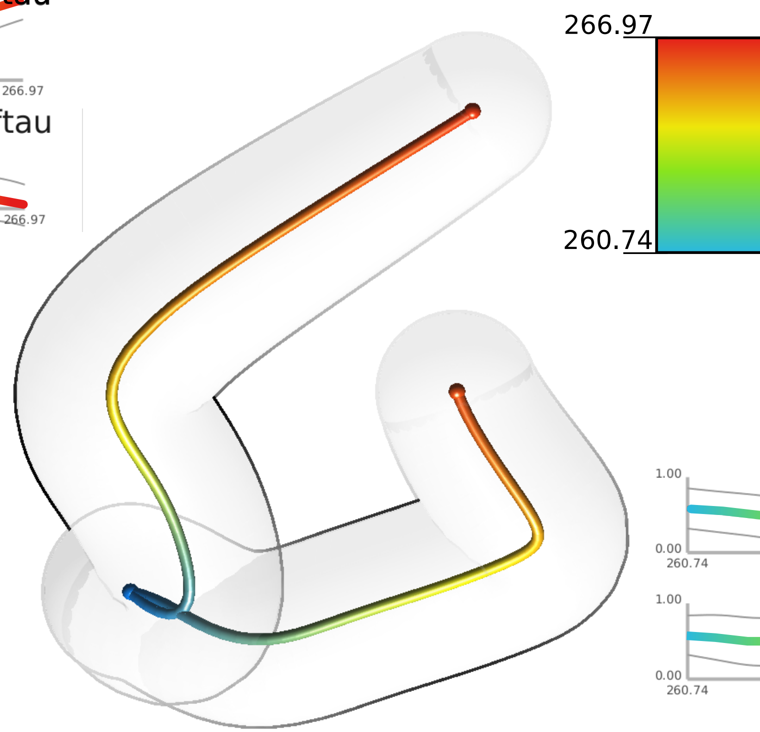


Topological Analysis Reveals Some Interesting Relationship Between Shallow/Deep Convection and Global Long Wave Flux

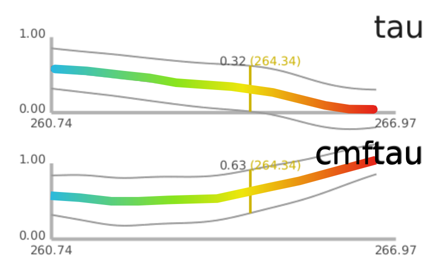
- The clear sky global long wave flux is maximal in two distinct parameter regimes something not evident through standard analysis



Inverse regression of tau/cmftau wrt. cell 1

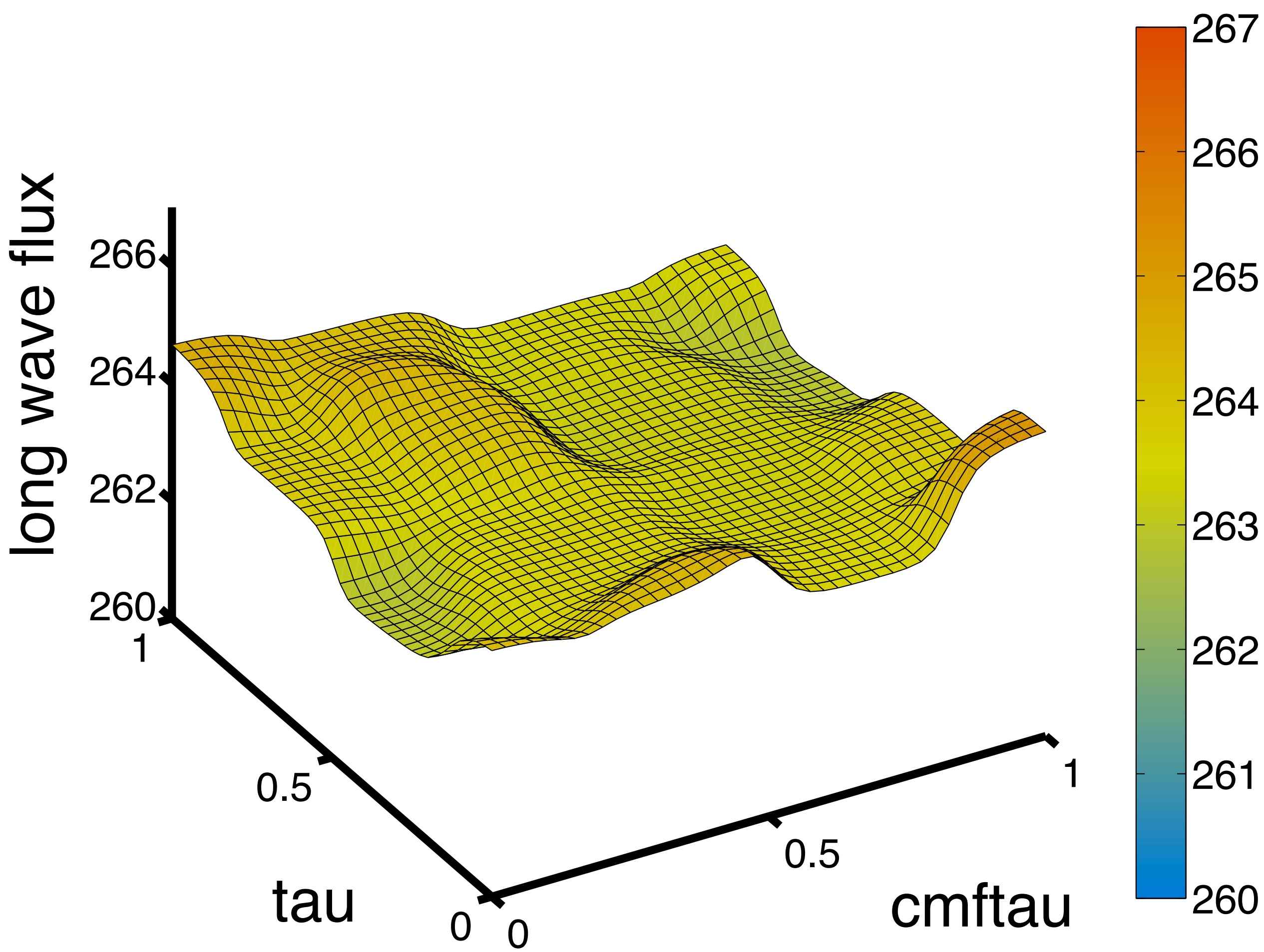


topology of FLUTC projected into 2D



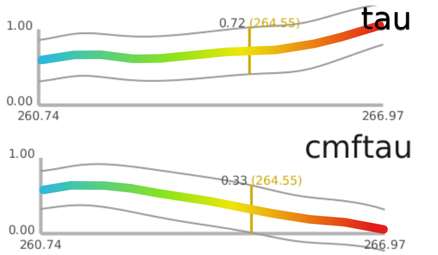
Inverse regression of tau/cmftau wrt. cell 2

The Same Information is Not Apparent in Either BIC Selection or Standard Regression

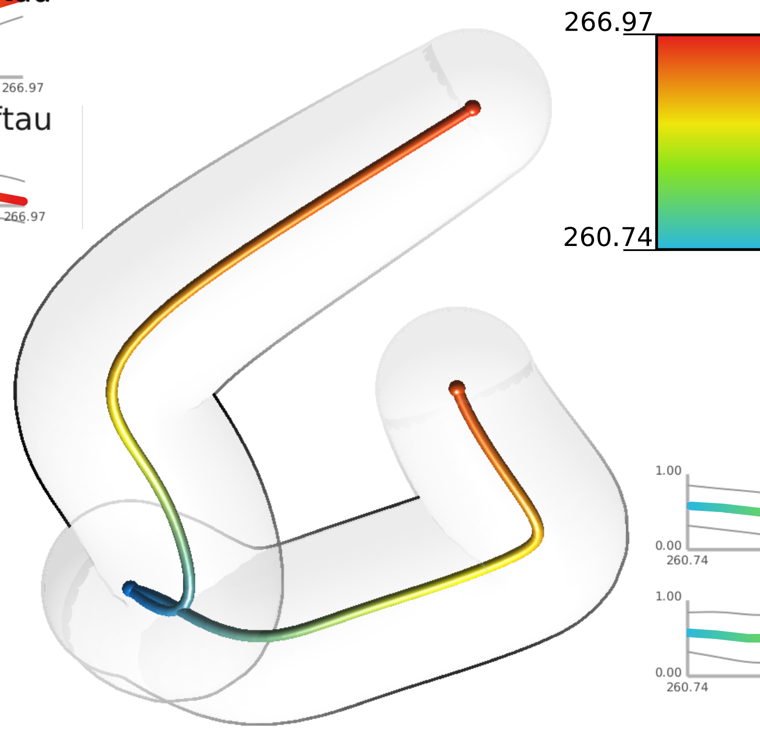


Topological Analysis Reveals Some Interesting Relationship Between Shallow/Deep Convection and Global Long Wave Flux

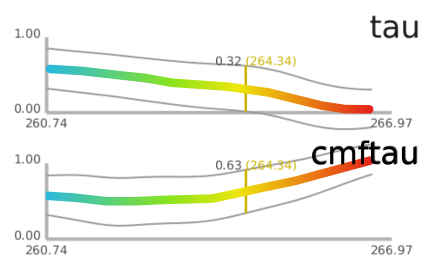
- The clear sky global long wave flux is maximal in two distinct parameter regimes something not evident through standard analysis



Inverse regression of tau/cmftau wrt. cell 1



topology of FLUTC projected into 2D



Inverse regression of tau/cmftau wrt. cell 2