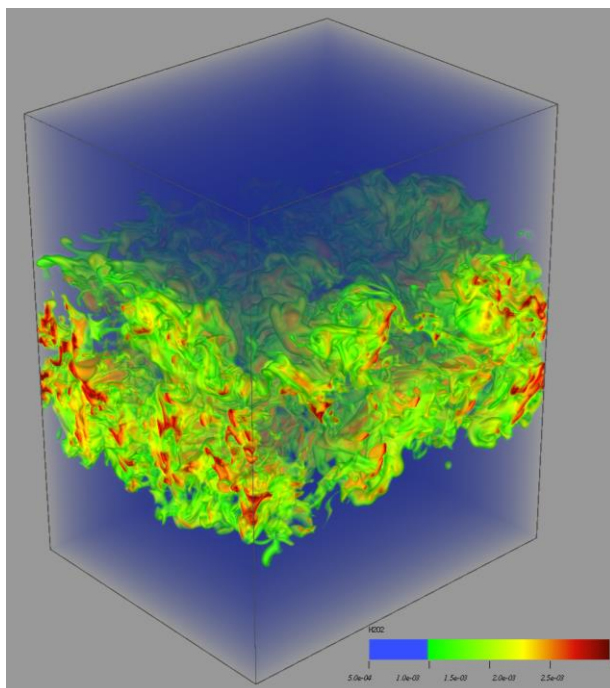
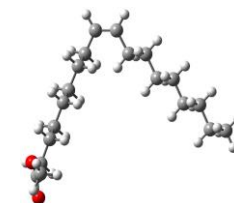
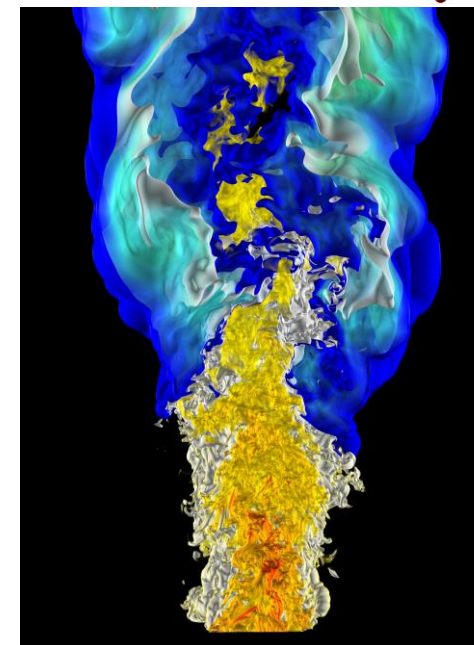


Framework for *In situ* Reduced Order Surrogate Models for DNS of Turbulent Reacting Flows at the Exascale



Jacqueline Chen
Senior Scientist
Sandia National Laboratories

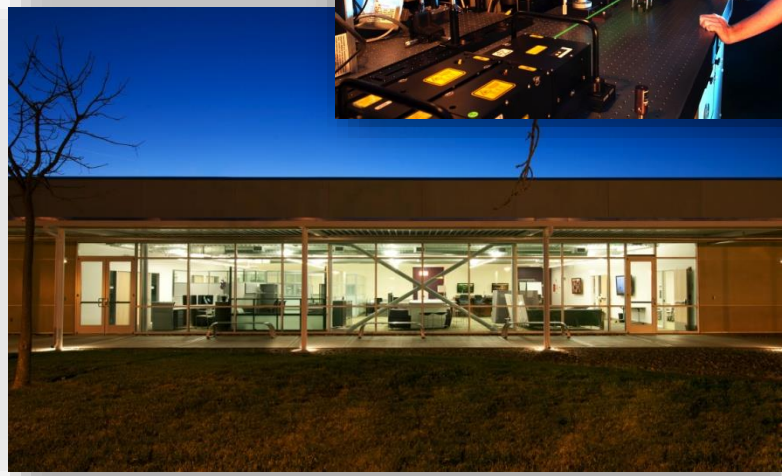
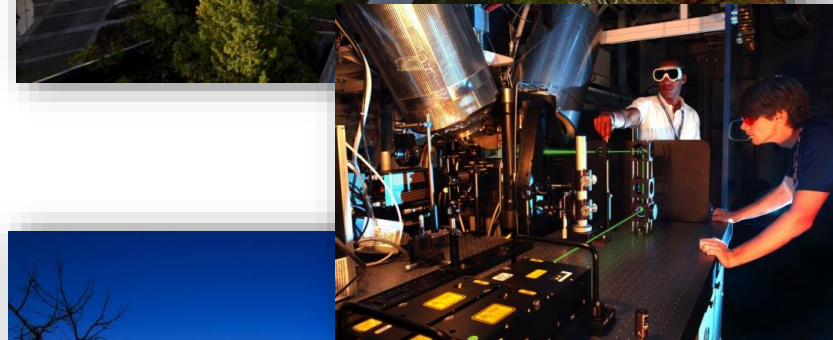
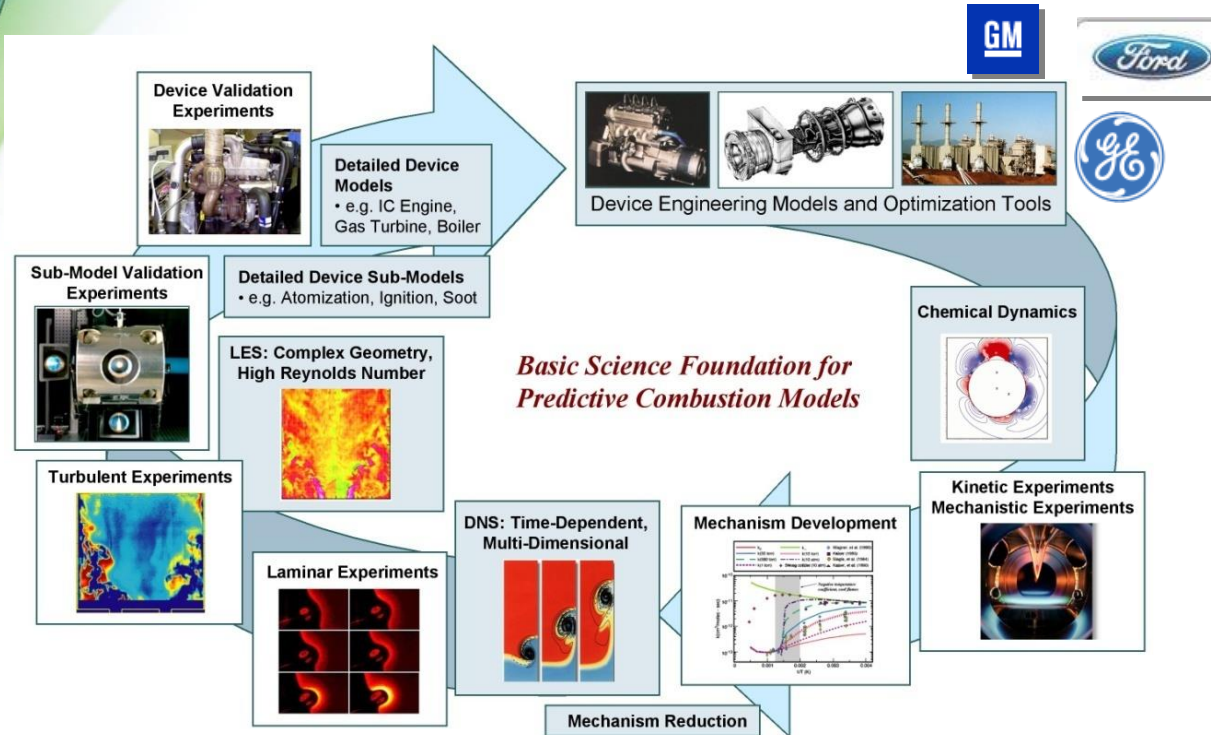
**SC Distinguished Scientists
Fellow 2020
ASCAC Meeting
September 25, 2020**





Sandia Combustion Research Facility

A DOE/BES Collaborative Research Facility dedicated to energy science and technology for the twenty-first century



Keys to CRF's success:

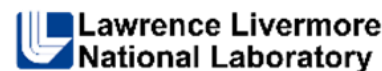
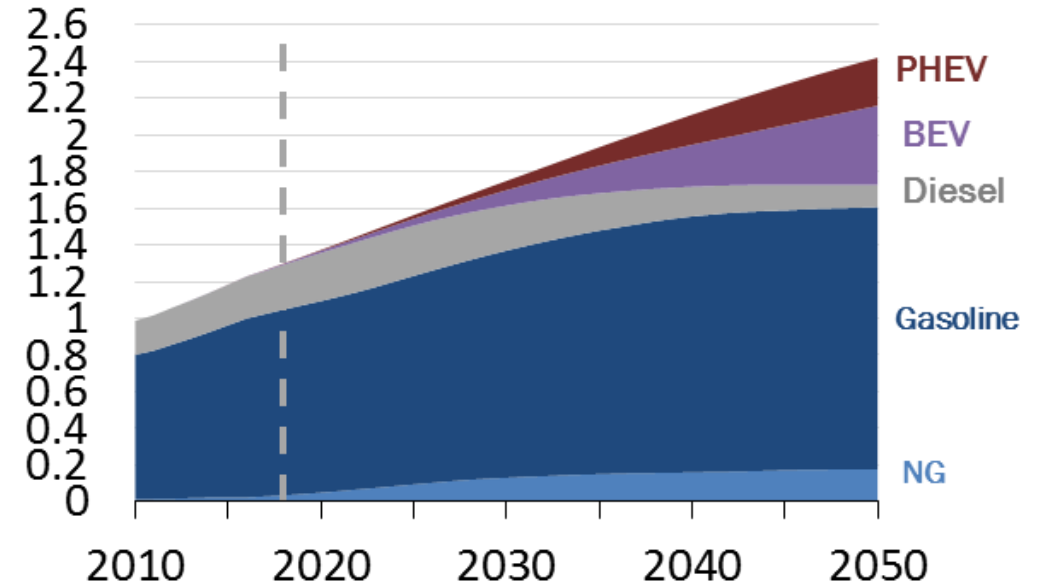
- Common scientific purpose
- Full spectrum of basic to applied
- Collocation and collaboration
- Strong ties to application and energy impact
- Visitor program adds intellectual vitality

Transportation will depend on internal combustion engines for decades

- Despite the expected growth in electric vehicle sales, the global light-duty fleet will be dominated by vehicles with internal combustion (IC) engines in the coming decades
 - Improving IC engine efficiency is an important path to energy security and reducing greenhouse gas emissions in the 2050 timeframe
- DOE Vehicle Technology Office Partnership in Advanced Combustion Engines (PACE) focuses resources across six national laboratories on common, key barriers to engine efficiency
- Leverages DOE investments in high performance computing through ECP and recent advances in ML/AI
- Direct path to industry OEM computational fluid dynamics workflow
 - Connect SC 'big science' to industry design process

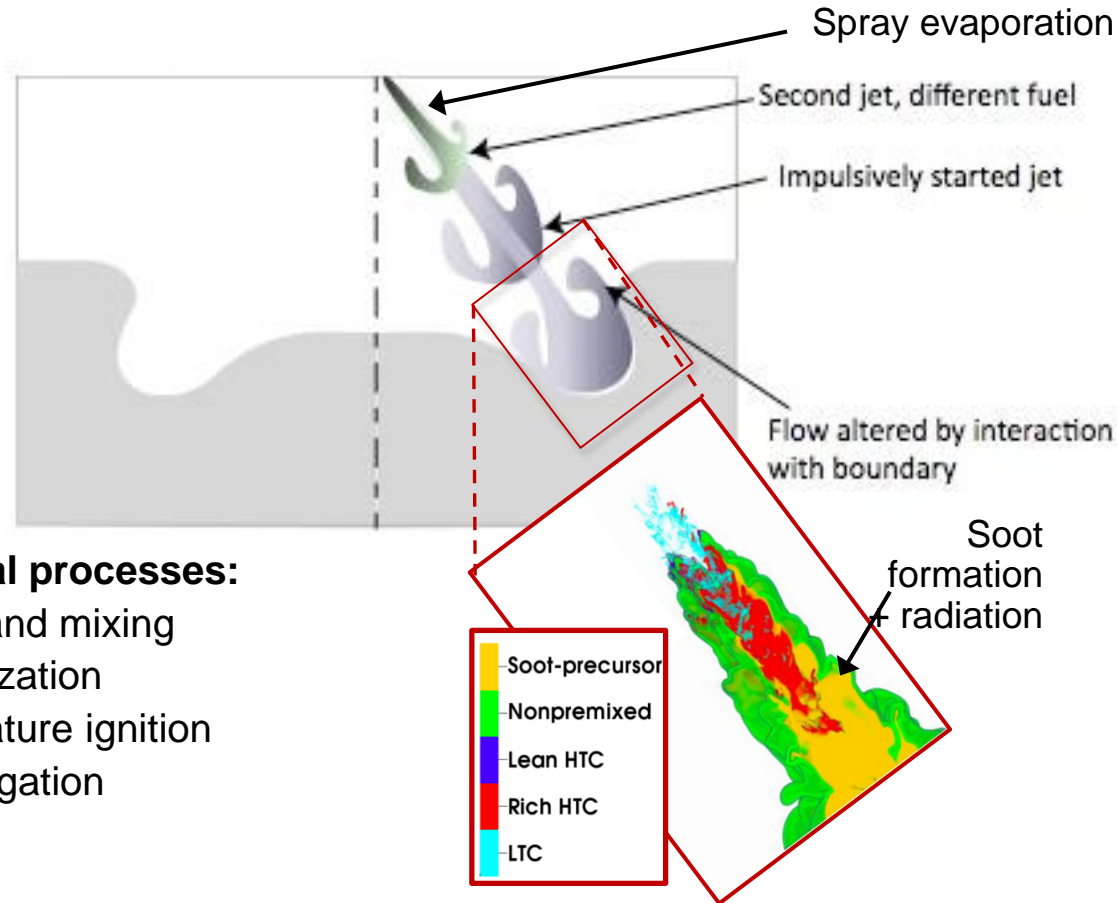
2019 International Energy Outlook, EIA

Light-duty vehicle stock
billion vehicles



ECP Pele Combustion High-Fidelity DNS Codes at the Exascale

Pele: reacting flow PDE solvers featuring block-structured adaptive mesh refinement with multi-physics for scalable performance portability on Frontier and Aurora in 2023



Critical physical processes:

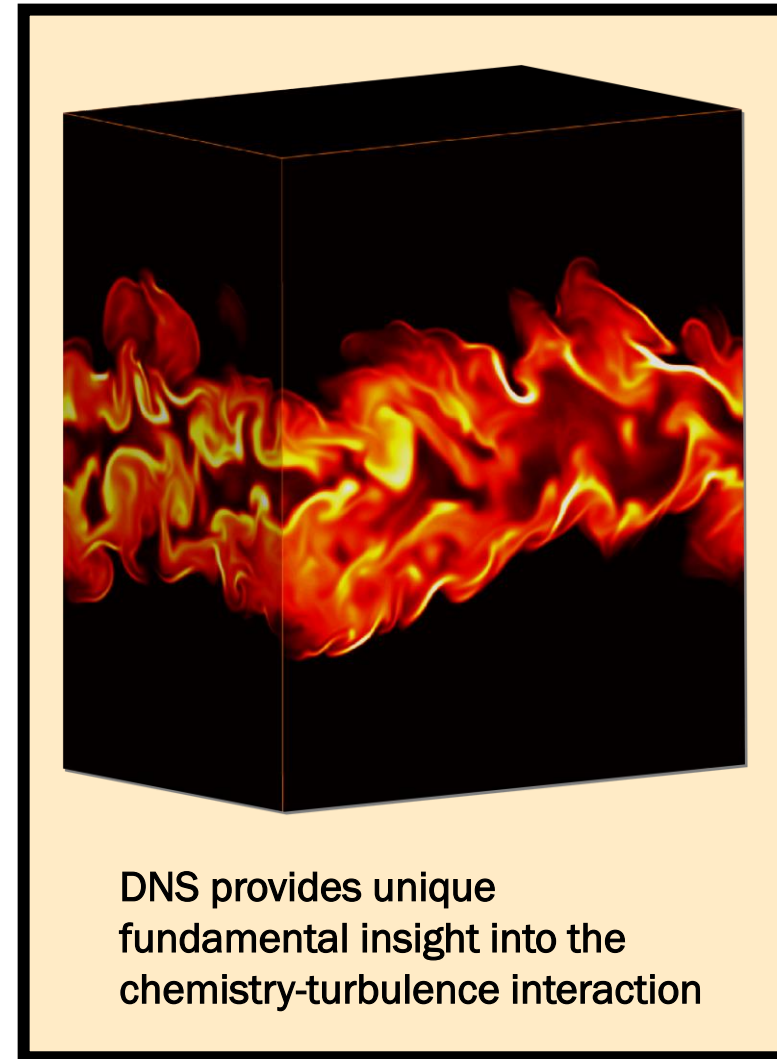
- ✓ Turbulence and mixing
- ✓ Spray vaporization
- ✓ Low-temperature ignition
- ✓ Flame propagation
- ✓ Soot
- ✓ Radiation
- ✓ Chemical Kinetics (leveraging BES Gas Phase Chemical Physics and automated combustion chemistry for drop-in mechanisms in CFD)

Key questions and sensitivities

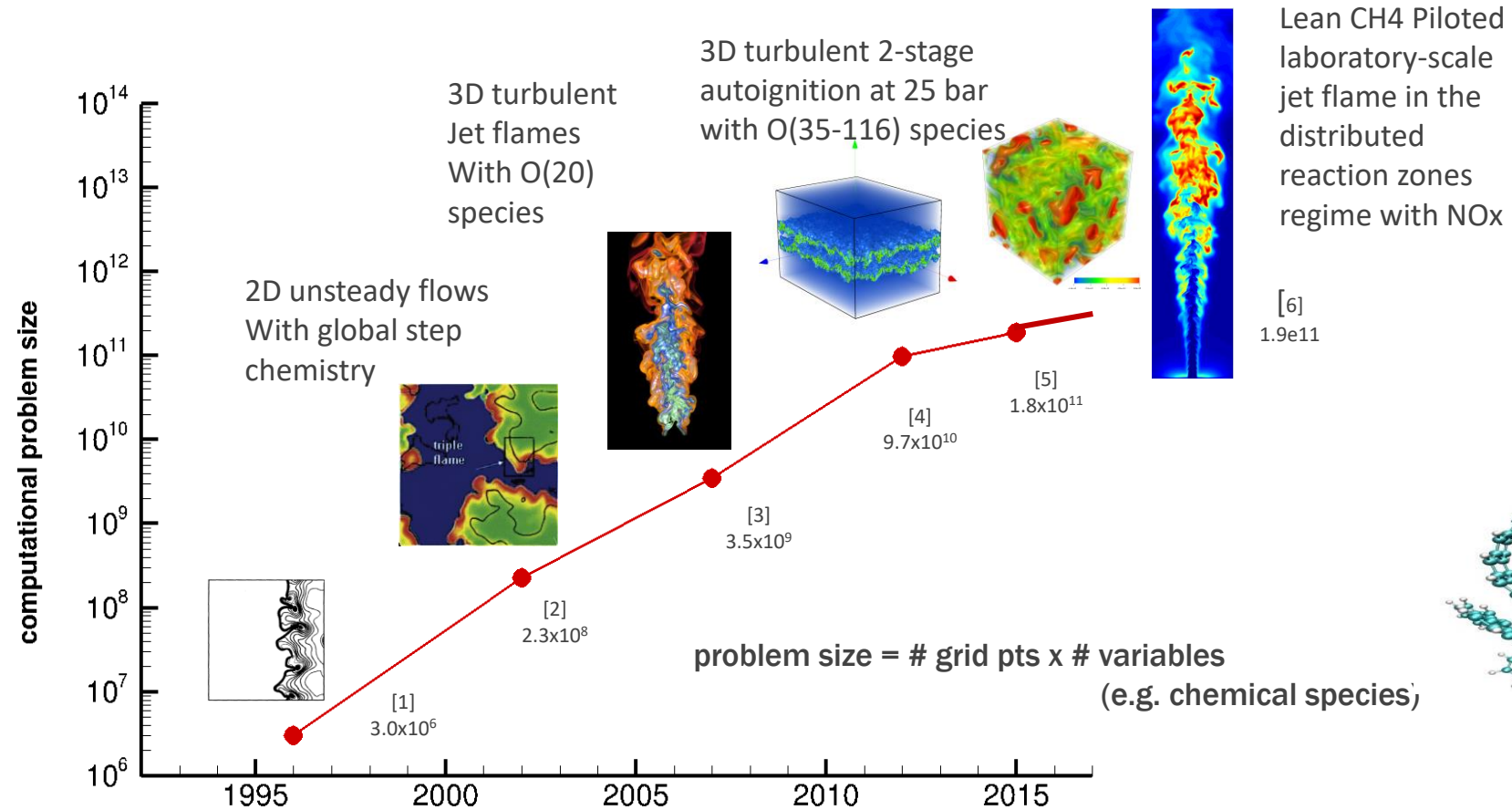
- What is the *distribution of reactivity* in a realistic mixture prepared by multiple injections
- How are *reactivity gradients* affected by injection characteristics:
 - Composition and mixing
 - Duration and timing
- How does the reactivity distribution and re-entrainment affect **soot generation**
- What are the *rate-controlling reactions* that expand the operating map for compression ignition engine combustion

Direct Numerical Simulation of Turbulent Reactive Flows – S3D

- Solves compressible reacting Navier-Stokes, total energy and species continuity equations
- High-order finite-difference methods
- Detailed reaction kinetics and molecular transport models
- Lagrangian particle tracking (tracers, spray, soot)
- *In situ* analytics and visualization
- Geometry using immersed boundary method
- Refactored for heterogeneous architectures using dynamic task based programming model (Legion)



Computational intensity of DNS scales with Moore's Law



[1] T. Echekki, J.H. Chen, *Comb. Flame*, 1996, vol.106.

[2] T. Echekki, J.H. Chen, *Proc. Comb. Inst.*, 2002, vol. 29.

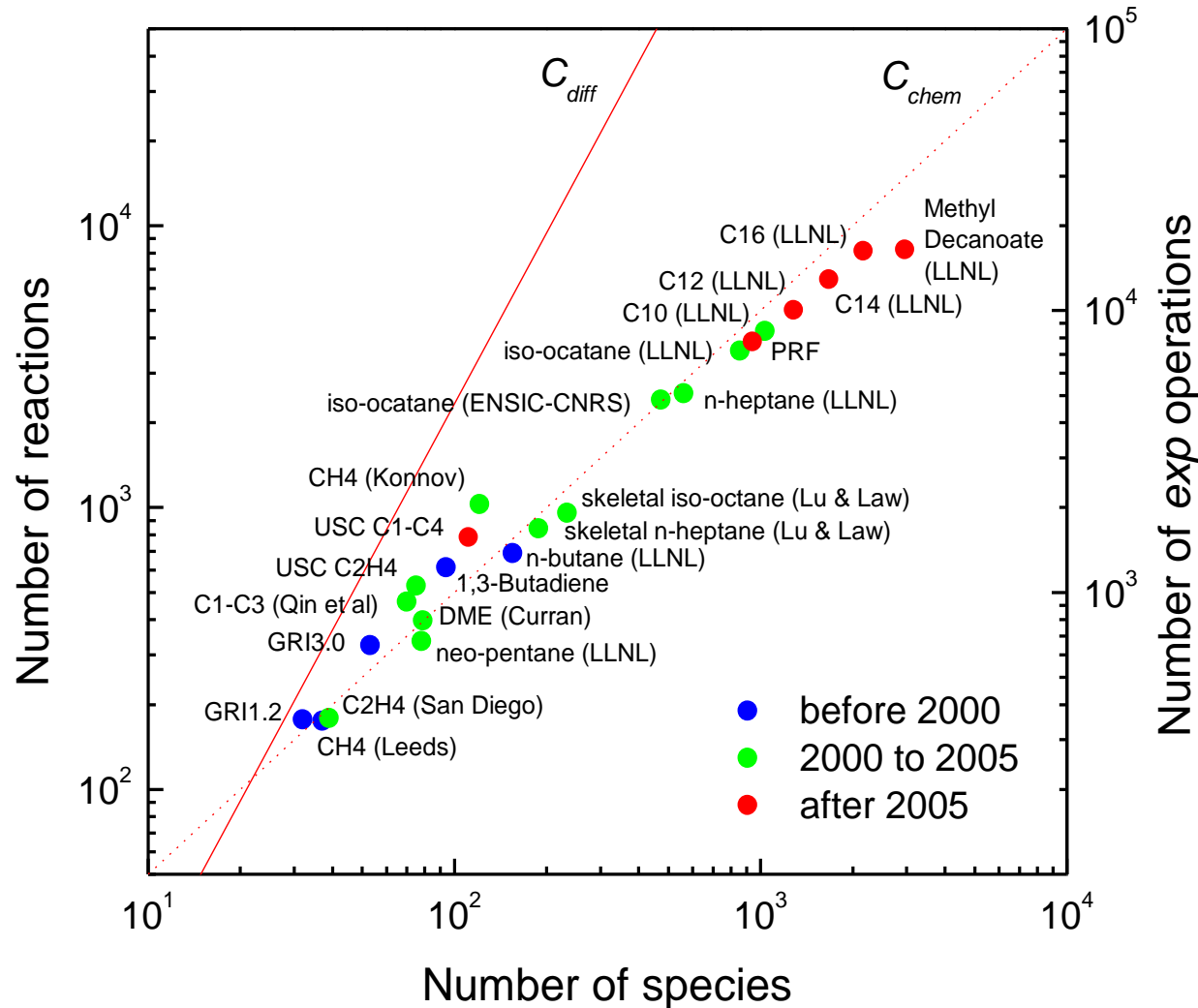
[3] R. Sankaran, E.R. Hawkes, J.H. Chen, *Proc. Comb. Inst.*, 2007, vol. 31.

[4] E.R. Hawkes, O. Chatakonda, H. Kolla, A.R. Kerstein, J.H. Chen, *Comb. Flame*, 2012, vol. 159.

[5] 2015 submission for Gordon Bell prize

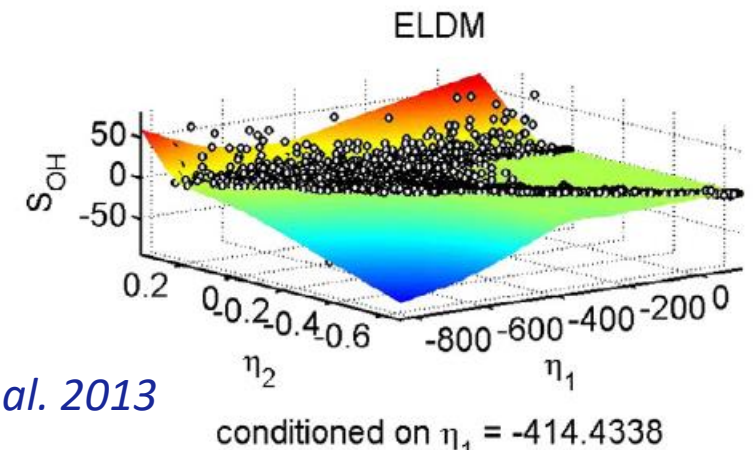
[6] H. Wang, E. Hawkes, J. H. Chen, *Comb. Flame* 2017

'Real' Fuels are described by chemical models with high-dimensionality: $O(1000)$ species and $O(10,000)$ reactions



Lu et al. 2009

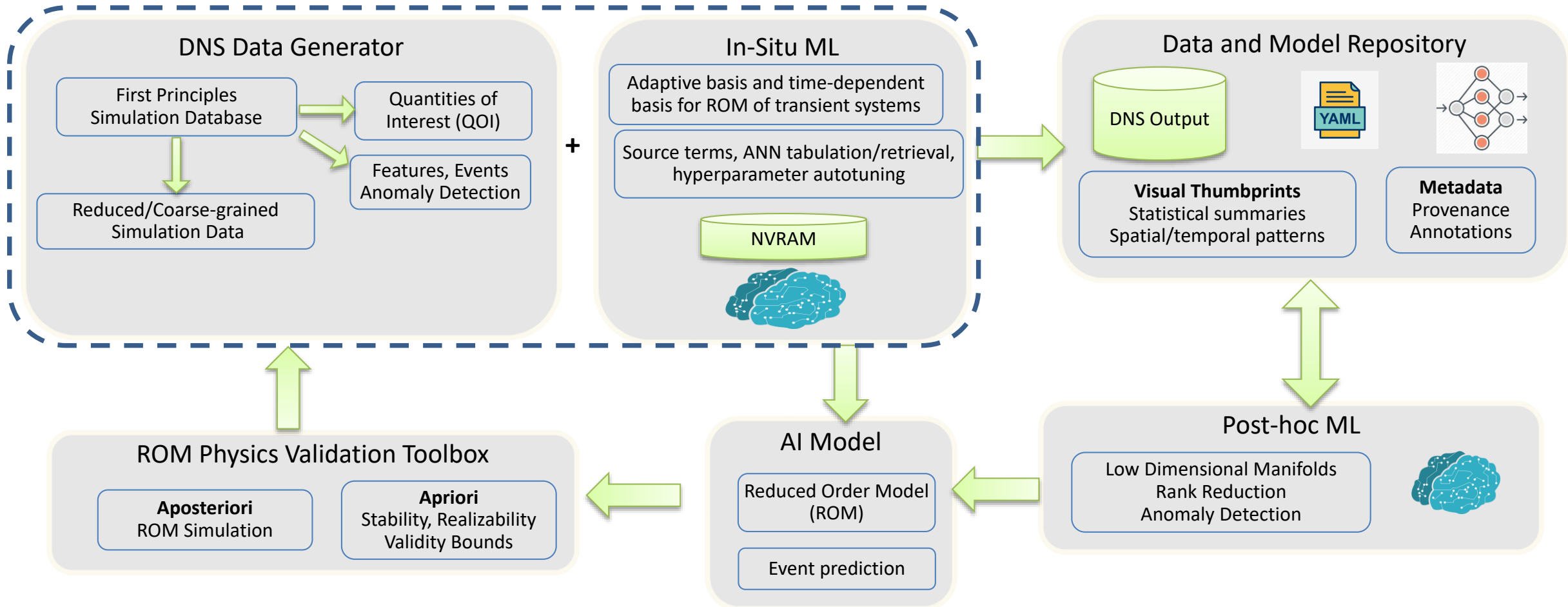
- Find low dimensional manifold in composition space that is a surrogate for full system dynamics of the reacting flow (to reduce the number of species transport equations to be solved)
- Manifolds exist due to inherent correlations of species (source of traditional paradigms for modeling)
- Empirical low-dimensional manifolds (ELDM) constructed from samples of compositions from experiment or DNS
- Linear (PCA) or nonlinear regression



Yang et al. 2013

Framework for *In situ* Reduced Order Surrogate Models for DNS of Turbulent Reacting Flows at the Exascale

In situ on Summit and on future exascale machines



Promote collaboration between DOE Labs and universities

Surrogate DNS with PC-Transport



Tarek Echehki
Professor
Mechanical & Aerospace Engineering
North Carolina State University



Infrastructure for complex DNS workflows using Legion/Regent/ FlexFlow with S3D



Hessam Babae
Assistant Professor
Mechanical Engineering
University of Pittsburgh



Michael Donello
MEMS PhD Student
(Fall 2018-Present)
University of Pittsburgh

In situ Sensitivity and ROMs in S3D



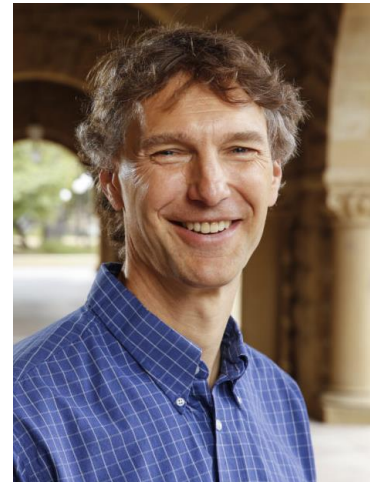
Yukiko Shimizu
Postdocs, Sandia National Laboratories



Martin Rieth



Elliott Slaughter
Associate Staff Scientist
Computer Science
SLAC

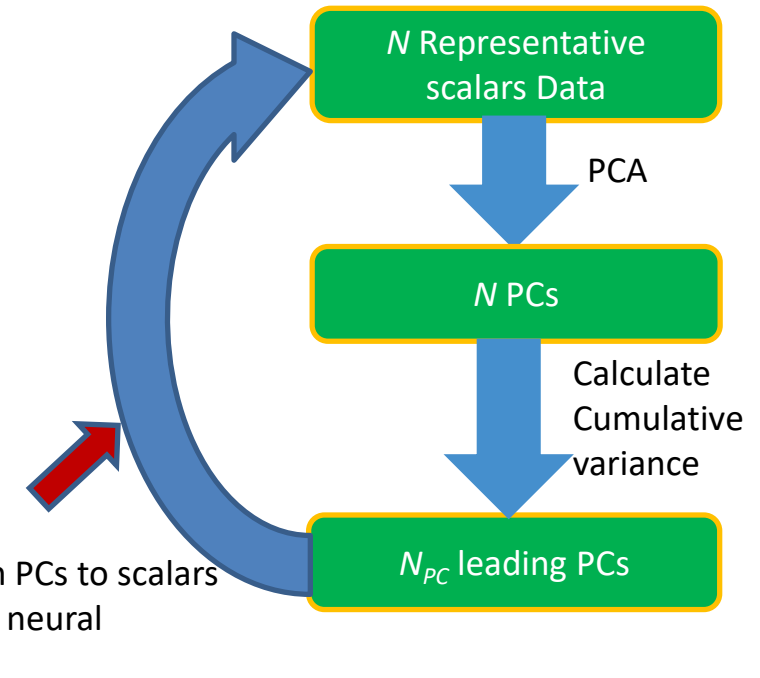


Alex Aiken
Professor
Computer Science
Stanford University

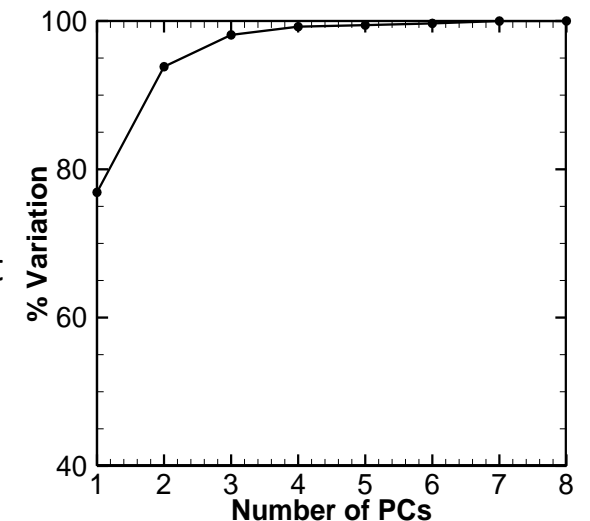
Real-time reduced order modeling of sensitivities & data compression in transient deterministic/stochastic systems, theory and algorithms

Surrogate DNS via *In situ* Adaptive Principal Component Transport

- Use ML/AI to aggressively reduce the high-dimensional composition space needed to describe gasoline and diesel surrogates, enabling DNS of ‘turbulence-chemistry’ interactions in SI gasoline and diesel engines
- Replicate high-dimensional composition space with low-dimensions using principle component (PC) transport analogous to species transport equations coupled with deep neural networks (DNNs) to model the chemical and transport terms (chemical source terms and diffusion coefficients) for PCs in terms of the transported PCs and to recover the original thermo-chemical scalars from the retained PCs.
- A few PC’s, linear combination of species compositions, can represent the variance of the original DNS with 1000’s of species - *potential for 2 orders of magnitude savings in DNS cost and storage*
- Instantaneous transport equations for the PCs in DNS can be derived (Sutherland and Parente, 2009):
- Chemical and transport terms in the PC-transport equations are similar to the species equivalent terms and are modelled using deep neural networks that relate them to the transported PC’s. Similarly, the original thermochemical scalar can be reconstructed from the PC’s.



Scree Plot



$$\frac{\partial \rho \phi_k}{\partial t} + \frac{\partial \rho u_j \phi_k}{\partial x_j} = \frac{\partial}{\partial x_j} \left[\rho D_k \frac{\partial \phi_k}{\partial x_j} \right] + s_{\phi_k}, \quad k = 1, \dots, N, \text{ where } s_{\phi} = A^T s_{\theta}$$

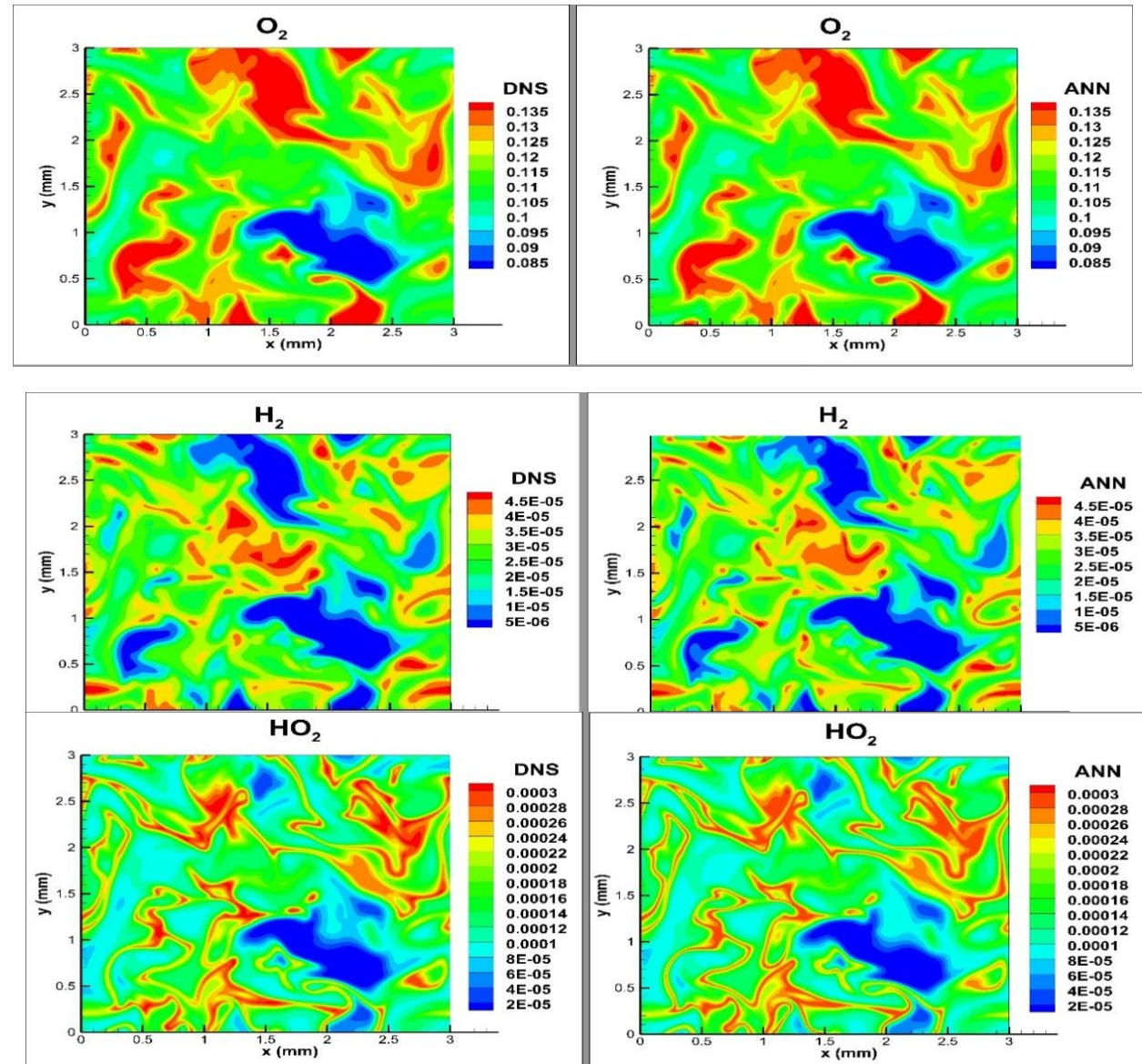
where A^T is a matrix with the leading eigenvectors of the covariance matrix

A Priori: DNS of Ethanol Compression Ignition

A Priori ... (an important step in PC transport)

- Determine PCs from slices of DNS solution
- Retain subset of the PCs
- Reconstruct solutions at different snapshots using the retained PCs.

- Original Manifold: **29 dimensional** (28 species + temperature)
- Linear PCA performed in a subspace spanned of **6 representative scalars**: T , O_2 , H_2O , CO , CO_2 and C_2H_5OH .
- PCA-ANN tabulation of all 29 variables are satisfactory based on first **2 PCs**.
- Potentially, an order of magnitude saving in computational time.
- Also potential for reducing stiffness if fast reactions are eliminated from the reduction process.



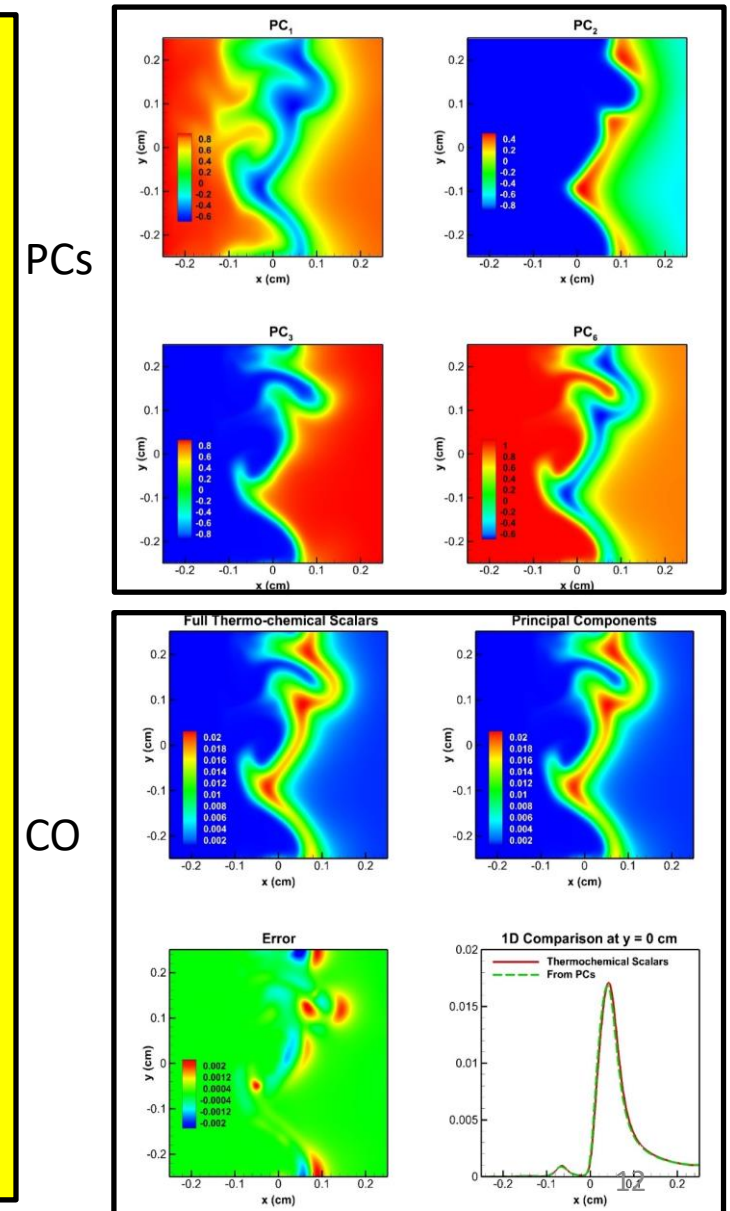
DNS data

PCA-ANN tabulation

A *Posteriori*: 2D DNS of methane-air premixed flame wrinkling (Owoyele & Echekki, 2018)

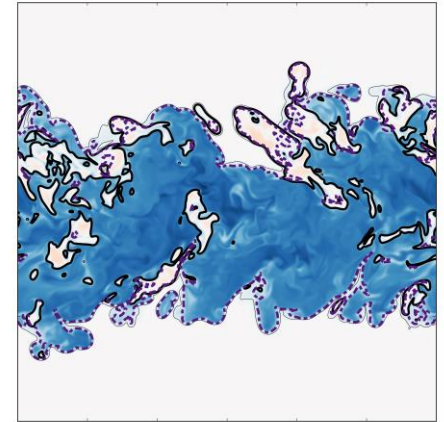
A *Posteriori*: PC Transport in DNS (an *a priori* step is needed prior to a *posteriori*)

- (31) Thermo-chemical scalars: (30 species + temperature) and 184 reactions
- (8) Representative scalars: T , CH_4 , O_2 , H_2O , CO_2 , CO , H_2 and O (O need to capture curvature/differential diffusion effects).
 - (8) Potential PCs
- (4) Retained PCs
 - The PCs capture the flame topology and are correlated with different key scalars.
- Saving in computational cost:
 - 4 vs. 31 scalars transported
 - A factor of 4 spatial resolution saving
 - A factor of 10 temporal resolution saving
- 2D DNS with species and energy has a similar computational cost to 3D DNS with PCs (huge saving)
- PC transport is not limited to a particular combustion mode.

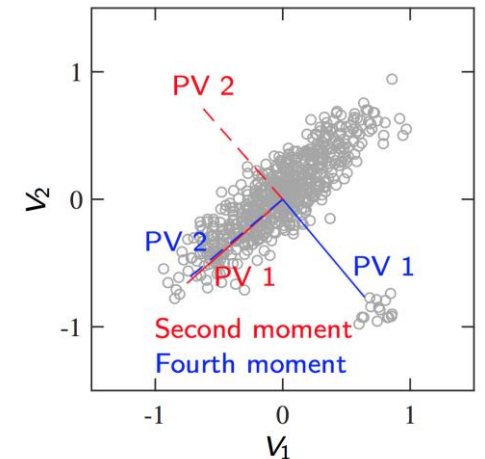


Next Steps with PC Transport

- Current state-of-the-art:
 - Start with static PCA from existing data generated from a low-dimensional simulation (2D vs. 3D) or smaller domain.
 - Requires knowledge of the composition space accessed by the simulations *a priori*
- PC transport can be optimized by dynamically evolving PCs during a simulation
 - Data on which PCA is carried out evolves with the simulation
 - Reduction and modeling of the transport terms for the PCs must be done on the fly during the simulation
 - Need to develop criteria for when PCs are dynamically updated
 - Need to develop strategies for transitioning from the old PCs to the new PCs
- Further extensions
 - PCA is designed to model the dominant features of a combustion problem
 - PCs will be augmented with additional bases that capture anomalies using anomaly detection algorithms
 - Anomaly detection provides criteria for transitioning the PC bases as well as to track the occurrence of rare events (e.g. extinction or ignition)

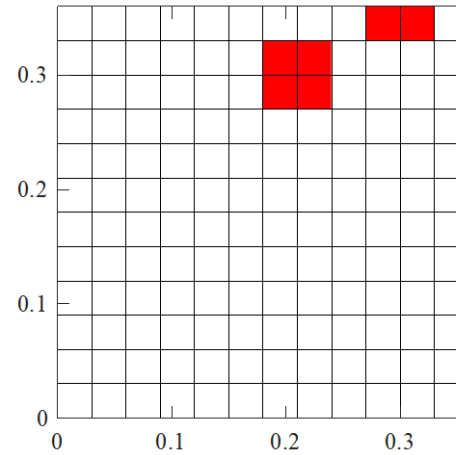
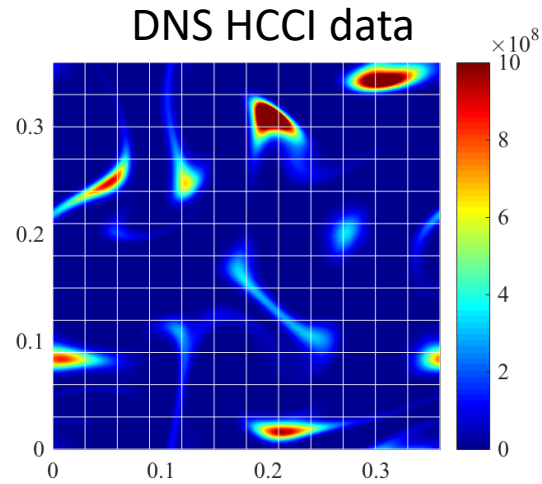


Temperature anomalies during autoignition

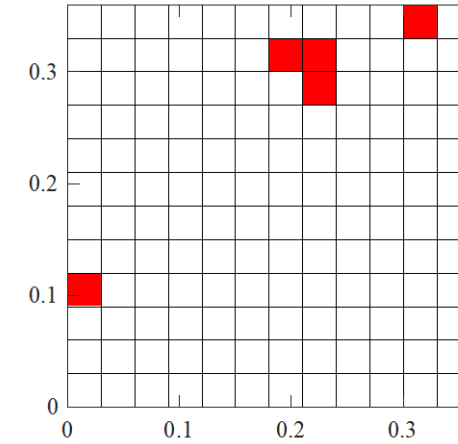


First principal kurtosis vector aligns in the direction of anomalies

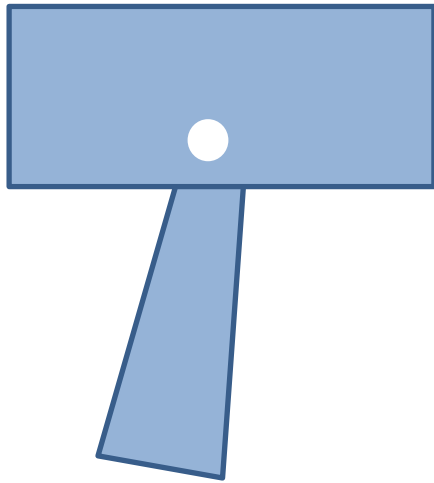
Anomaly detection using joint statistical moments as a trigger for adapting PC-transport subspace (Example: predict pre-ignition knock and its indicators for GDI SI engines at boosted conditions)



True ignition kernels



Autoignition predicted by anomaly detection



- Anomaly detection to identify ignition regions (trigger for adaptive PC-transport)
- An anomaly detection algorithm [1] using factorization of higher moment tensor (co-kurtosis) was used to identify spatio-temporal regions where auto-ignition was occurring in a DNS simulation of HCCI combustion. Plot on left shows contour map of heat release rate at an instant of low temperature ignition, the center shows the sub-domains containing the true ignition kernels, the right plot shows sub-domains identified by the anomaly detection algorithm.
- Implement in-situ light-weight anomaly detection ML tasks on accelerators

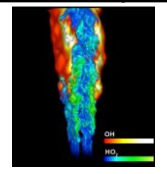
Real-Time Reduced Order Modeling Via Time-Dependent Basis: *Model-Driven and Data-Driven Modalities*

◆ **Optimality:** Find the time-dependent subspace $\mathbf{U}(x, t)$ and its coefficients $\mathbf{Y}(t)$ by minimizing the functional

$$\mathcal{F}(\dot{\mathbf{U}}(x, t), \dot{\mathbf{Y}}(t)) = \left\| \frac{d(\mathbf{U}(x, t)\mathbf{Y}(t)^T)}{dt} - \mathcal{M} \right\|_F^2 \quad \dot{(\cdot)} = \frac{d}{dt} \longrightarrow \begin{cases} \dot{\mathbf{U}} = f_U(\mathbf{U}, \mathbf{Y}) \\ \dot{\mathbf{Y}} = f_Y(\mathbf{U}, \mathbf{Y}) \end{cases}$$

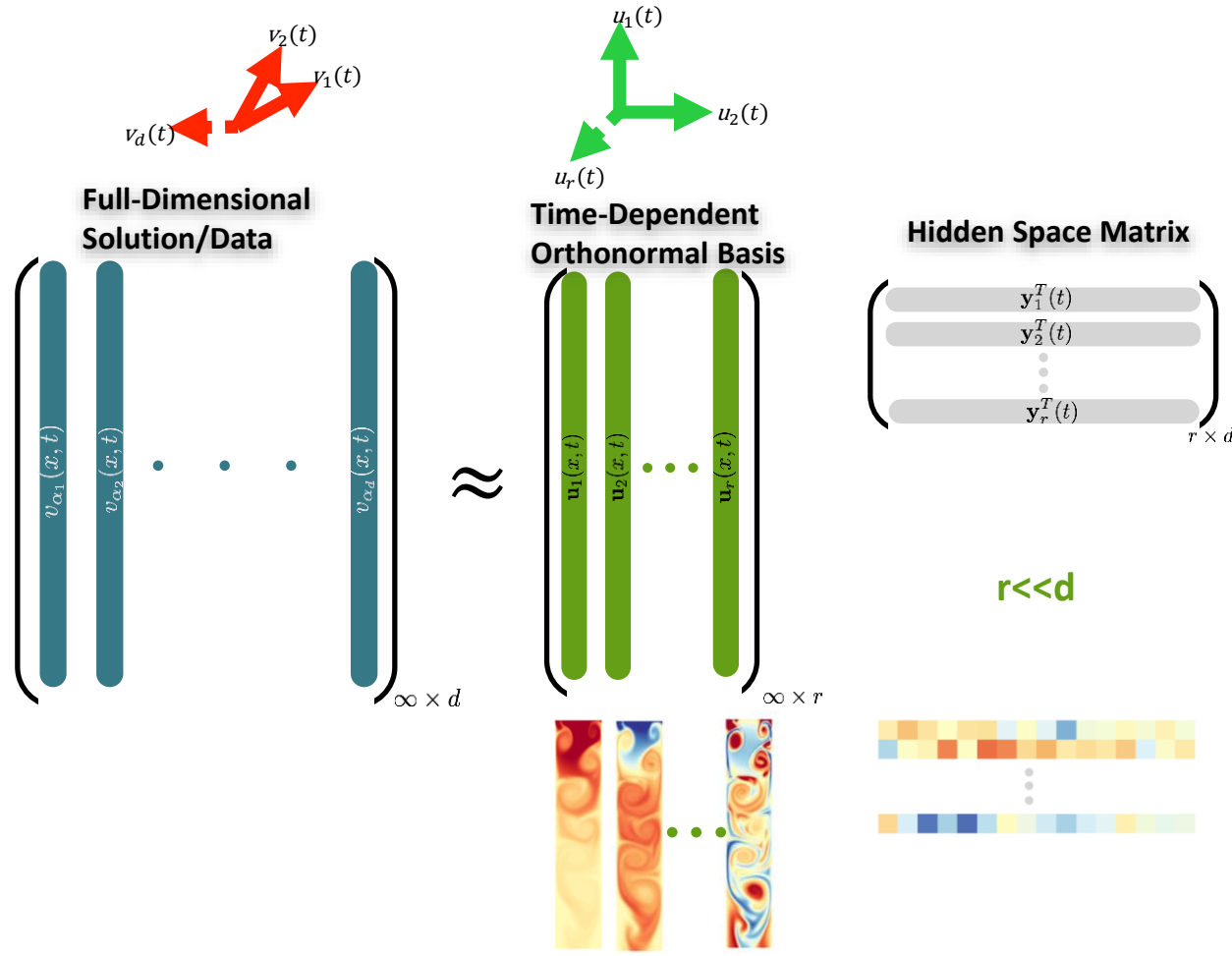
Model (PDE) or Data (from simulation)

◆ **On-the-fly Data/Model Compression:** Extracts low-rank structure from streaming simulation **data** or directly from **model**. Application: **in-situ** compression.

Data-Driven Non-intrusive	<p>Data: Mass Fraction (DNS)</p>  <p>Streaming simulation data</p>	<p>In-situ Low-Rank Extraction</p> <p>Advance $\mathbf{U}(x, t)$ & $\mathbf{Y}(t)$ to the next time step</p> $\begin{cases} \dot{\mathbf{U}} = f_U(\mathbf{U}, \mathbf{Y}) \\ \dot{\mathbf{Y}} = f_Y(\mathbf{U}, \mathbf{Y}) \end{cases}$	<p>Compression/ROM</p> <ul style="list-style-type: none"> ◆ Build real-time reduced order model. ◆ Store $\mathbf{U}(x, t)$ & $\mathbf{Y}(t)$ instead of the full solution.
Model-Driven Intrusive	<p>Model: Sensitivity Equation (Large Number of Parameters)</p> <p>Solve the sensitivity equation in the compressed form.</p> <p>PDE for $\mathbf{U}(x, t)$ $\dot{\mathbf{U}} = f_U(\mathbf{U}, \mathbf{Y})$</p> <p>ODE for $\mathbf{Y}(t)$ $\dot{\mathbf{Y}} = f_Y(\mathbf{U}, \mathbf{Y})$</p>		

◆ **Knowledge Discovery:** Time-dependent basis $\mathbf{U}(x, t)$ discovers low-dimensional subspace of systems with finite-time instabilities/rare events (e.g. ignition/extinction/blowoff/flashback, turbulent intermittency)

◆ **Scalable Method:** Scales linearly with respect to size of data and low-rank r . Does not require solving large scale eigenvalue/optimization problem.



Extract spatial/parametric correlations from model or data on the fly

Journal of Computational Physics 415 (2020) 109511

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Real-time reduced-order modeling of stochastic partial differential equations via time-dependent subspaces

Purna Patil, Hessam Babaei*

Department of Mechanical Engineering, University of Pittsburgh, Pittsburgh, PA-15260, United States of America

PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Babaei H. 2019 An observation-driven time-dependent basis for a reduced description of transient stochastic systems. *Proc. R. Soc. A* 475: 20190506. <http://dx.doi.org/10.1098/rspa.2019.0506>

An observation-driven time-dependent basis for a reduced description of transient stochastic systems

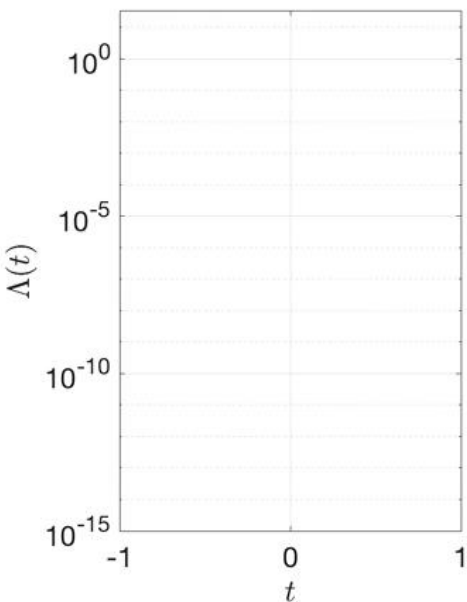
H. Babaei

Department of Mechanical Engineering and Materials Science, University of Pittsburgh, 3700 O'Hara Street, Pittsburgh, PA 15261, USA

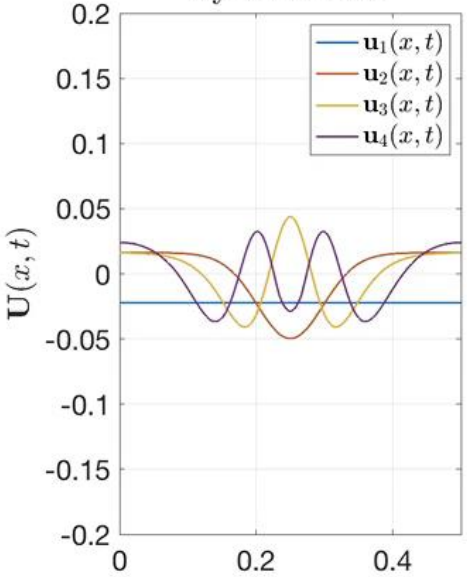
HB, 0000-0002-6318-2265

Data-Driven Reduction: *On-the-fly* Compression of DNS Combustion Data

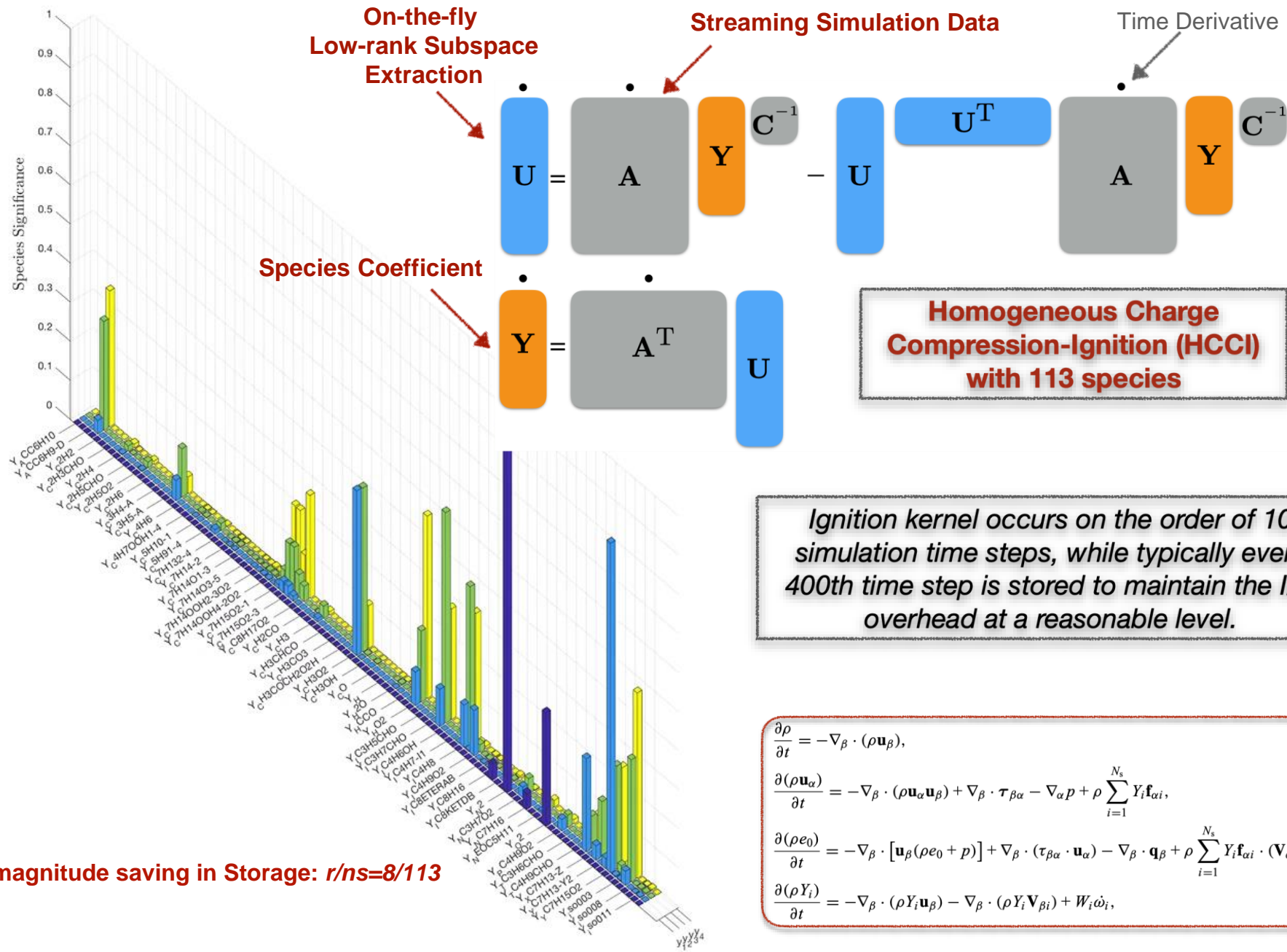
Covariance Matrix



Dynamic Basis

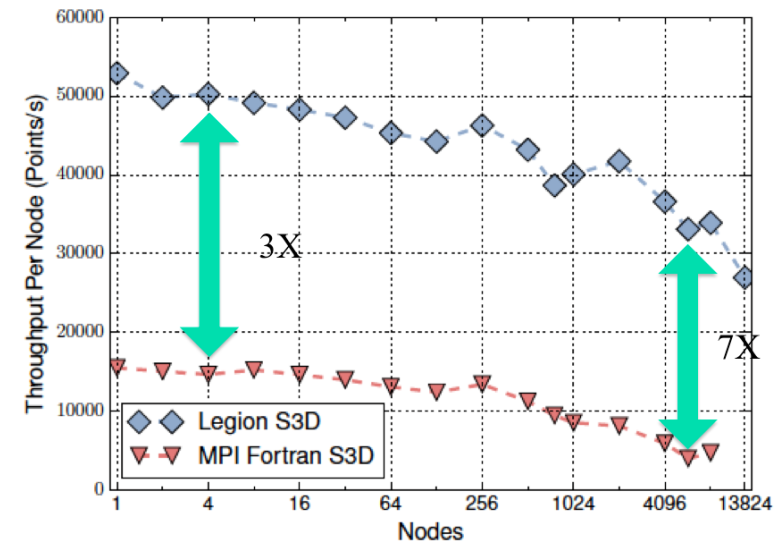
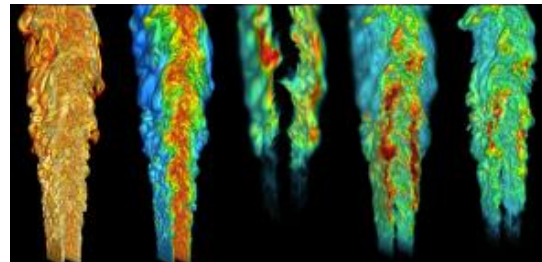


An order of magnitude saving in Storage: $r/ns=8/113$



Legion Programming System Applied to S3D

- A data-centric parallel programming system
- A programming model for **heterogeneous, distributed** machines
 - Automates many aspects of achieving high performance, such as extracting task- and data-level parallelism
 - Automates details of scheduling tasks and data movement (*performance optimization*)
 - Separates the specification of tasks and data from the mapping onto a machine (*performance portability*)
- Legion application example: S3D DNS
 - Production combustion simulation
 - Written in ~200K lines of Fortran
 - Direct numerical simulation using explicit methods



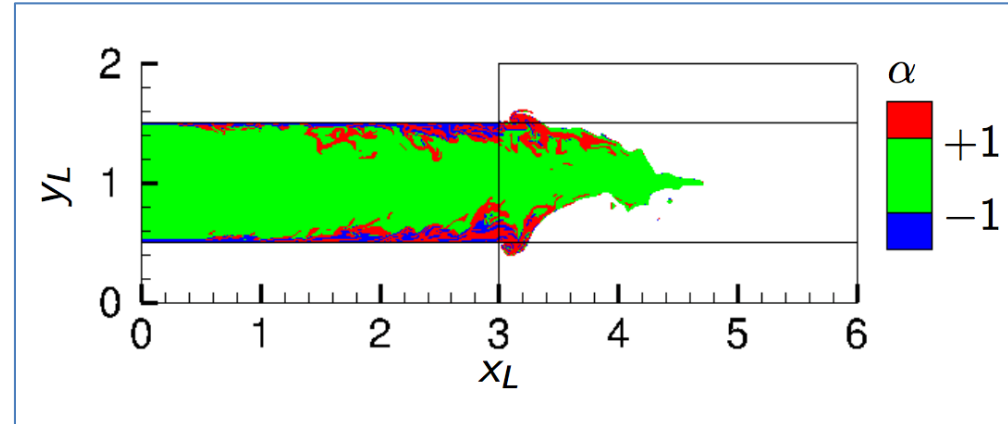
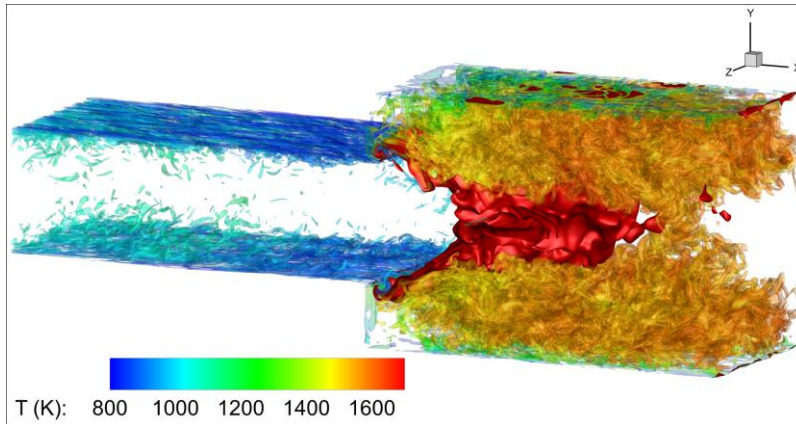
S3D performance Legion vs. MPI

S. Treichler et al., "S3D-Legion: An Exascale Software for Direct Numerical Simulation (DNS) of Turbulent Combustion with Complex Multicomponent Chemistry," CRC Book on Exascale Scientific Applications: Programming Approaches for Scalability Performance and Portability, 2017.

In-situ Data Analytics in Legion S3D

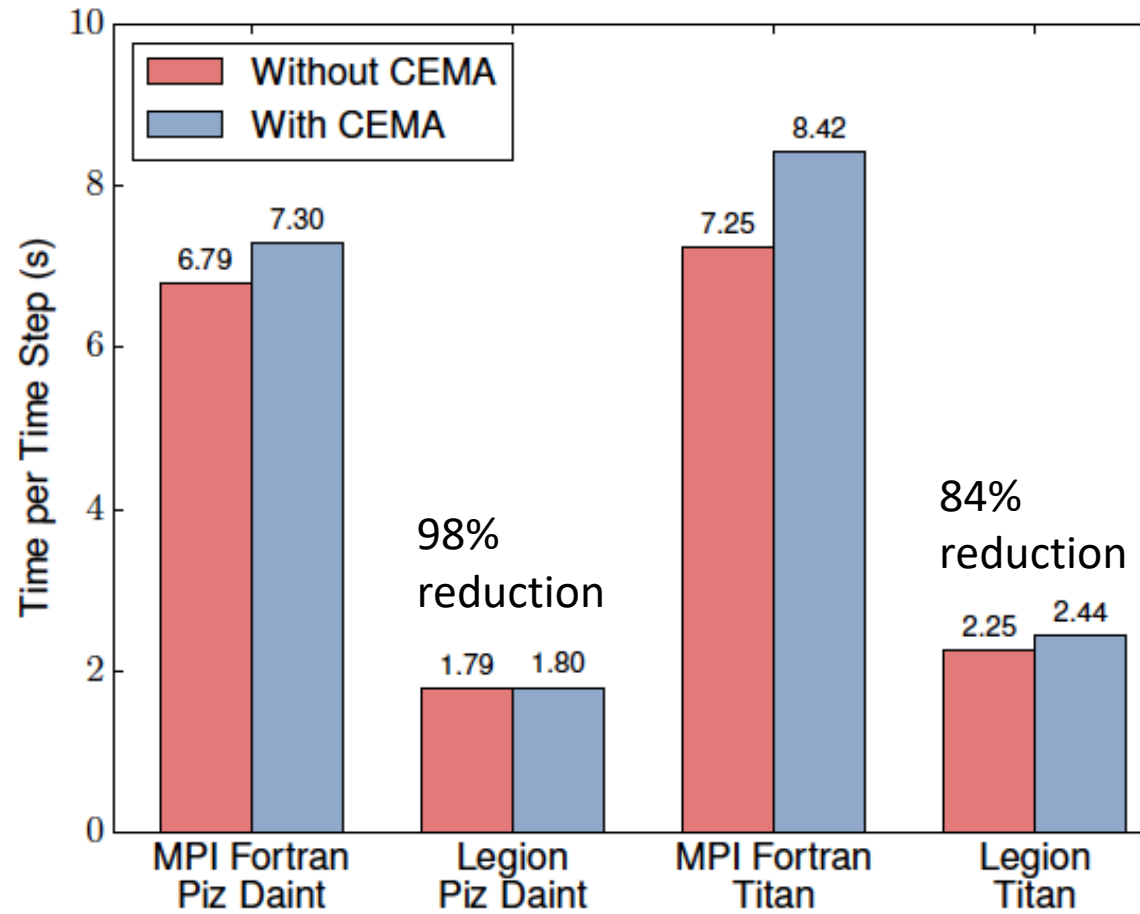
Chemical Explosive Mode Analytics (CEMA)

- CEMA: eigenvalue solve on the reaction rate Jacobian to determine the mode of combustion (ignition, flame propagation, extinction)



- Run CEMA every 10 time steps as a diagnostic for identifying combustion mode
- CEMA computation takes longer than a single explicit RK stage (6 stages/timestep)
- Dividing CEMA across RK stages and interleaving with other computation so as not to impact other critical operations would be hard to schedule manually
- Asynchronous task execution, schedule CEMA on CPU resources
- Interoperate Fortran CEMA with Legion S3D DNS code – took a day to implement

Execution Overhead of In-situ Analytics (CEMA) in Legion-S3D (Titan & Piz Daint)



Regent code allows application scientists to write code with sequential semantics

- Task-based programming model
 - Built on the Legion runtime
 - Used in ExaFEL, S3D, PSAAP II & III, ...
 - Ports to exascale machines in progress
 - Automatically compiles for different GPU's
- Key features
 - Transparent support for code on CPUs or GPUs
 - Expressive data partitioning for distributed computation
 - Compiler and runtime manage scheduling, communication, data placement, ...
 - Highly portable

```
__demand(__cuda)
```

```
task CalcVolumeTask(lr_q : region(ispace(int3d), fields.QFields),
```

```
lr_int : region(ispace(int3d), fields.IntFields))
```

```
where
```

```
reads(lr_q.{RHO, RHO_U, RHO_V, RHO_W}),
```

```
writes(lr_int.{VOLUME, VEL_X, VEL_Y, VEL_Z})
```

```
do
```

```
for idx in lr_q.ispace do
```

```
var rho = lr_q[idx].RHO
```

```
var volume:double = 1.0 / rho
```

```
lr_int[idx].VOLUME = volume
```

```
lr_int[idx].VEL_X = volume * lr_q[idx].RHO_U
```

```
lr_int[idx].VEL_Y = volume * lr_q[idx].RHO_V
```

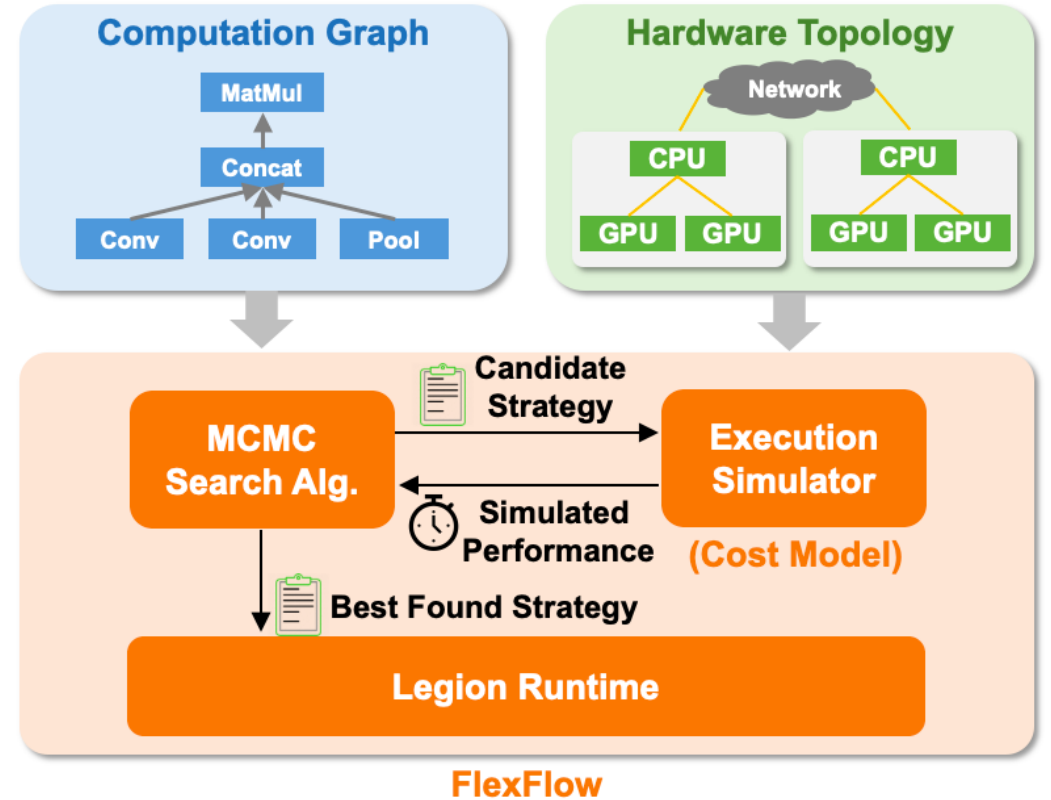
```
lr_int[idx].VEL_Z = volume * lr_q[idx].RHO_W
```

```
end
```

```
end
```


FlexFlow will be used for deep learning for parallel performance

- A distributed deep learning framework
 - Also built on Legion
 - Secret sauce: Automatic search to find a high-performance data partitioning
 - Dramatically improves locality and scalability
 - Reduces large scale training from days to hours
- Supports Keras interface
 - Pytorch support in progress



Promote scientific and academic excellence in ASCR and BES research through collaborations between Labs and universities

- **Co-Design:** collaboration between computer scientists, applied mathematicians (ML), and computational scientists at DOE Labs (SNL, SLAC) and universities (North Carolina State U., U. Pittsburgh, Stanford U.)
- **Train graduate students and postdocs** (extended visits at Labs), provide a pipeline for future computational and data scientists at DOE Labs
- **Develop an open source in situ DNS/ML framework** to evaluate ML/AI algorithms on heterogeneous exascale machines
- **Provide a scalable, portable DNS/ML framework** for composing complex workflows with PDE solver coupled with analytics/visualization/ML
- **Generate high-fidelity large-scale turbulent reactive flow simulation data** including metadata (provenance and annotations) for training and validation of ML/AI models
- **Engage with LCFs** on computing and infrastructure issues related to large-scale simulation of turbulent reactive flows with *in situ* data analytics and ML/AI

Promote collaboration between DOE Labs and universities

Surrogate DNS with PC-Transport



Tarek Echehki

*Professor
Mechanical & Aerospace Engineering
North Carolina State University*



Infrastructure for DNS/ML workflows using Legion/Regent/ FlexFlow with S3D



Hessam Babae

*Assistant Professor
Mechanical Engineering
University of Pittsburgh*



Michael Donello

*MEMS PhD Student
(Fall 2018-Present)
University of Pittsburgh*

In situ Sensitivity and ROMs in S3D



Yukiko Shimizu

Postdocs, Sandia National Laboratories

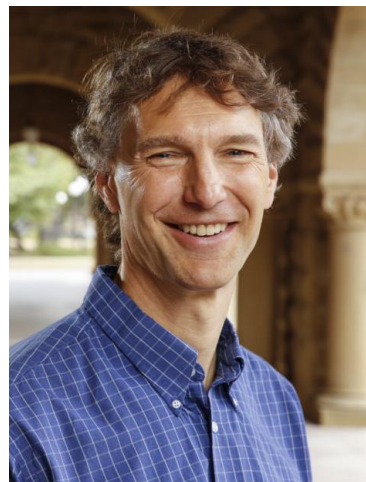


Martin Rieth



Elliott Slaughter

*Associate Staff Scientist
Computer Science
SLAC*



Alex Aiken

*Professor
Computer Science
Stanford University*

Real-time reduced order modeling of sensitivities & data compression in transient deterministic/stochastic systems, theory and algorithms