Advanced Modeling & Simulation (AMS) Seminar Series
NASA Ames Research Center, August 18th, 2022

Combustion Machine Learning: Principles, Progress, Perspective

MATTHIAS IHME, WAI TONG CHUNG AND AASHWIN MISHRA

mihme@stanford.edu

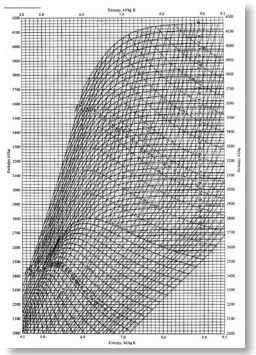


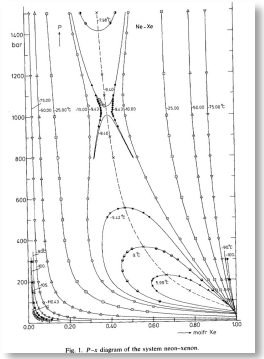


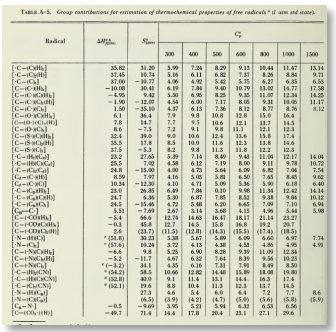


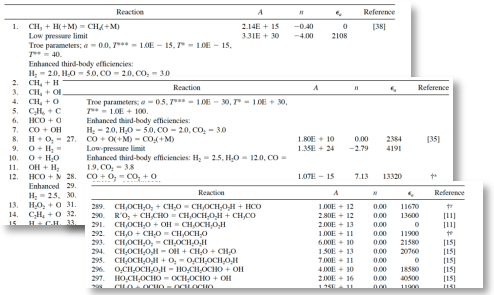
Data in combustion science and engineering

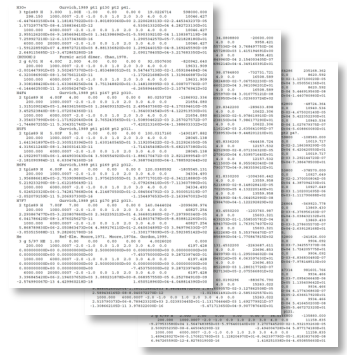
- Thermo-chemical properties
  - Formation enthalpies
  - Thermodynamic properties
  - Transport properties
  - Phase properties
  - Rate parameters









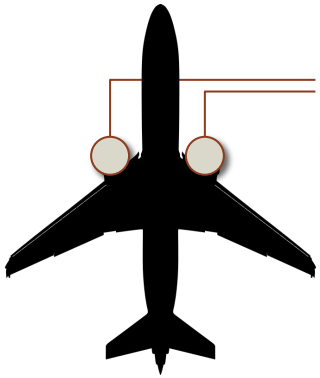


Fischer, Dryer, Curran, Int. J. Chem. Kinet. 32: 713–740, 2000
McBride, Zehe, Gordon, NASA/TP—2002-211556m 2002
Ruscic et al. J. Phys. Chem. A 108 (45), 9979-9997, 2004
NIST Chemistry Webbook

Stanford University

Data in combustion science and engineering

- Sensors and actuators
  - System prognostics
  - Health monitoring
  - Stall and flow control
  - Clearance
  - Emissions
  - Cooling



20 TB × 2 × 6 × 20,000

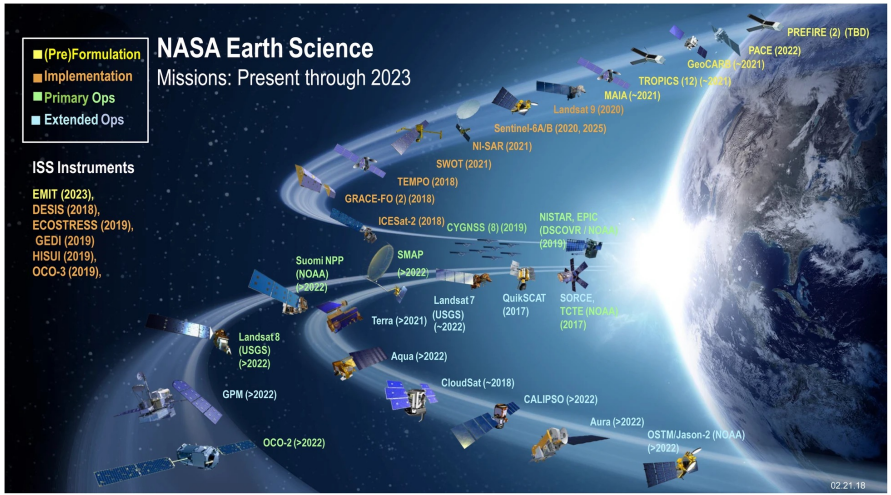
Data per engine and hour | Twin-engine Boeing 737 | 6h flight: JFK → LAX | USA daily commercial flights

= 4.8 EB/day

Garg et al., ASME GT2010-22685, 2010

Data in combustion science and engineering

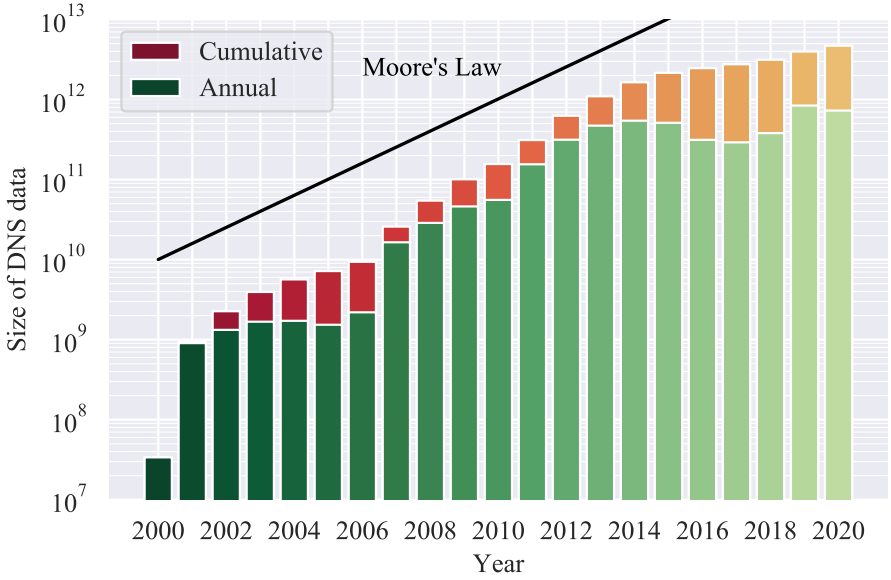
- Satellite observations
  - Wildfires
  - Vegetation, fuel density
  - Moisture, humidity
  - Emissions, pollutants
  - Smoke

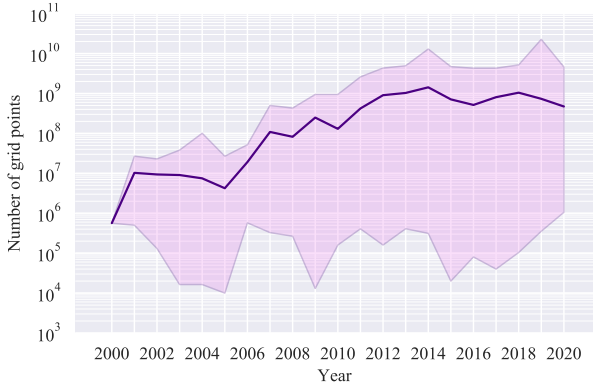


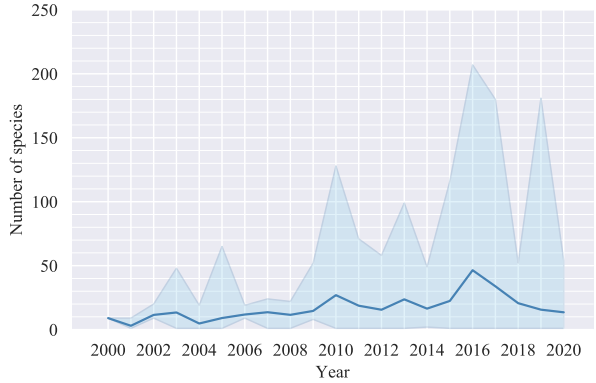
> 1TB/day/satellite

Ustin and Middleton, Ecol. Proc. 10, 2021.

Data in combustion science and engineering





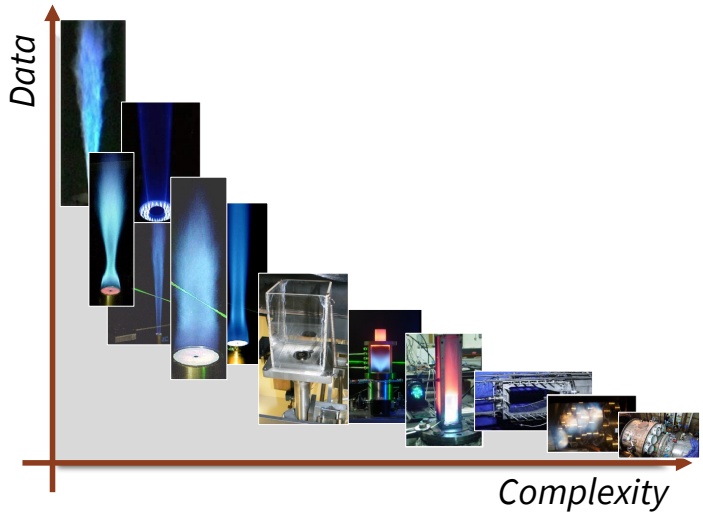


Ihme, Chung, and Mishra, Prog. Energy Combust. Sci. 91, 101010, 2022

Data in combustion science and engineering

Challenges

- Data limited to specific operating conditions, fuels, and geometries
- Lack of data for complex combustion conditions
- Data accessibility
- Prevailing domain knowledge



Data in combustion science and engineering

Challenges

- Well-established theoretical foundation of thermofluid flows
  - Thermodynamic principles
  - Conservation relations
  - Law of mass action
- Solution limited by
  - Wide range of scales
  - Chemical complexity
  - Incomplete knowledge of constitutive relations

Conservation equations (momentum, species, energy)

$$\underbrace{\partial_t U}_{} + \nabla \cdot \underbrace{F(U)}_{} - \nabla \cdot \underbrace{Q(U, \nabla U)}_{} = \underbrace{S(U)}_{}$$

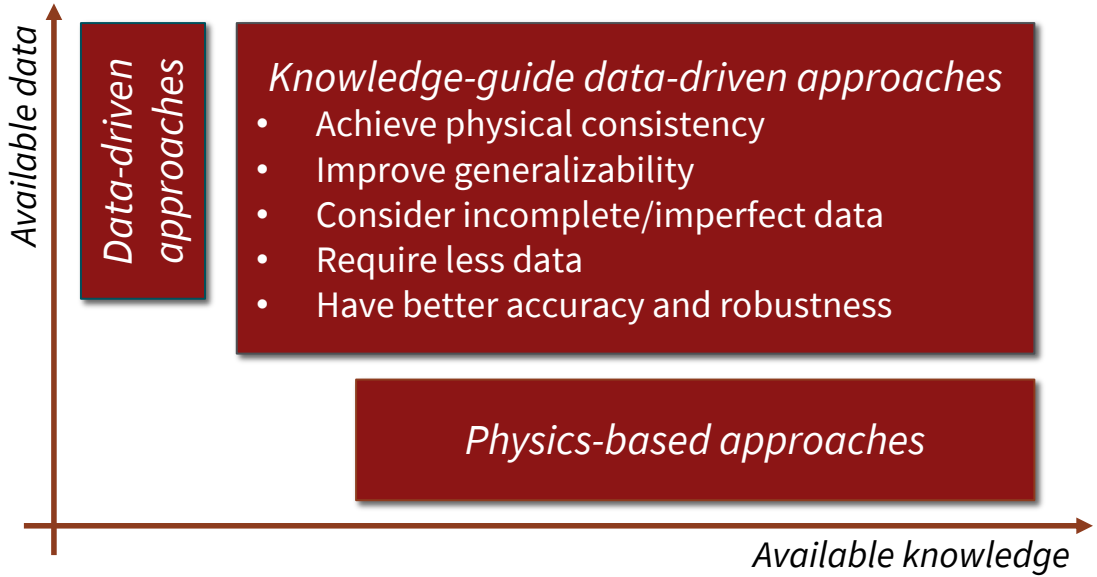
$$\begin{pmatrix} \rho u \\ C \\ \rho e_t \end{pmatrix} \quad \begin{pmatrix} \rho u \otimes u + p\mathrm{I} \\ C \otimes u \\ u(\rho e_t + p) \end{pmatrix} \quad \begin{pmatrix} \tau \\ -j \\ \tau \cdot u - q \end{pmatrix} \quad \begin{pmatrix} 0 \\ \dot{\omega} \\ 0 \end{pmatrix}$$

State relation $g(p, e, C) = 0$

Conservation and coupling functions

$$W^T C = \rho\,, \quad W^T \dot{\omega} = 0\,, \quad W^T j = 0$$

Knowledge discovery paradigms



Questions and overview

How can combustion science and engineering embrace ML advances?

- Review ML methods
- Leverage of knowledge and data
- Applications: Fundamental combustion science, propulsion and energy conversion, Hazards and safety management

What are unique challenges in applying ML to combustion?

- Data sets
- Integration of domain knowledge
- Uncertainty quantification and generalizability
- CombML-specific methods and techniques

ML Methods



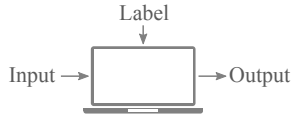
Machine learning methods for combustion

- *Machine learning:* enable computer-based systems to infer predictive models from data

- *Learning:* ability of an algorithm to progressively improve its performance in specific task by processing data and information (in absence of domain knowledge)

Machine learning methods for combustion

Supervised learning



Relate labeled input to unknown output

Unsupervised learning



Identify pattern and discover structures

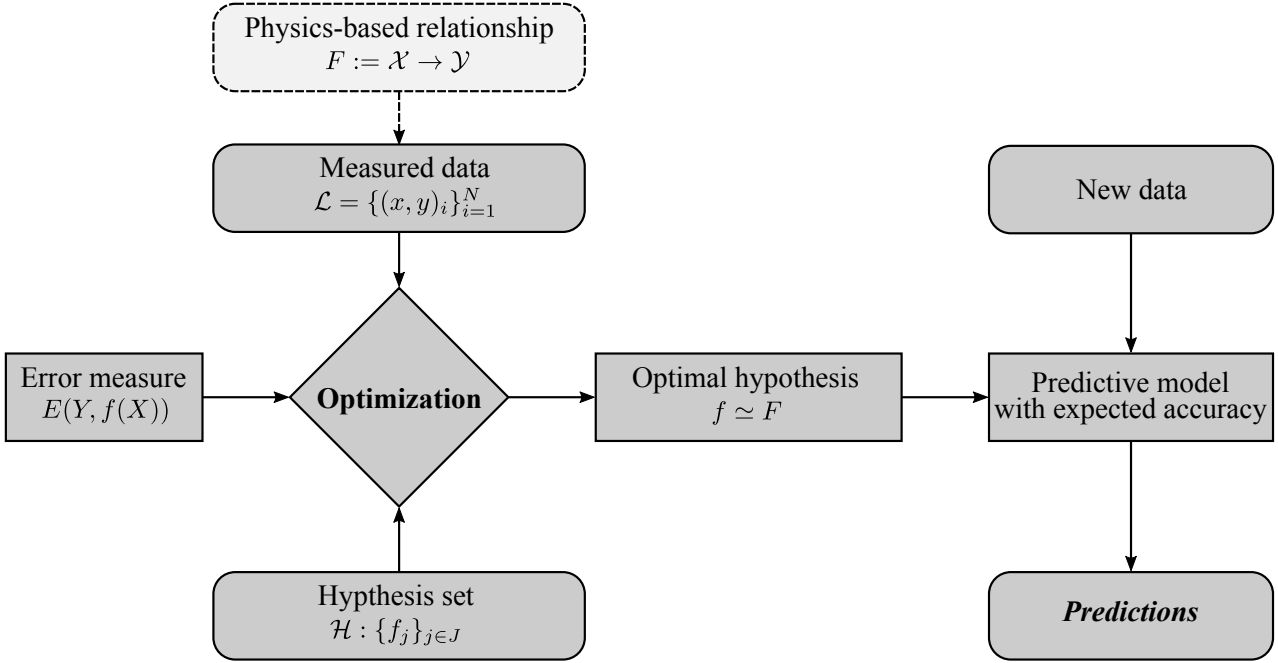
Semi-supervised learning



Learn from partially labeled data or through interaction with environment

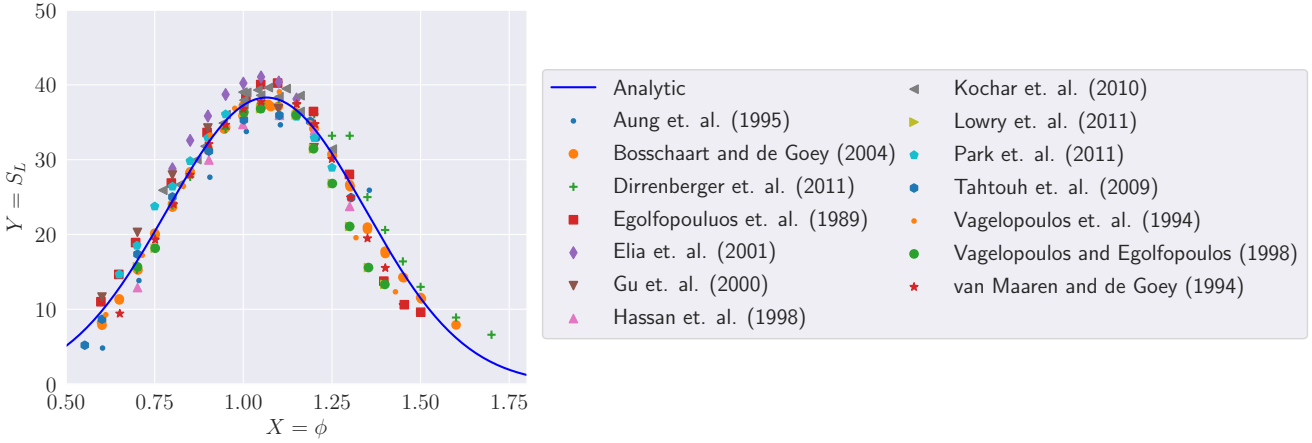
Machine learning methods for combustion

A generic supervised learning algorithm



Machine learning methods for combustion

Learning problem: polynomial approximation of flame-speed for CH4/Air mixture at ambient conditions

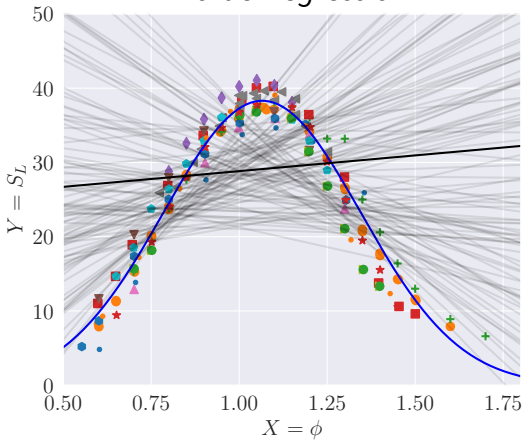


Machine learning methods for combustion

Learning problem: polynomial approximation of flame-speed for CH4/Air mixture at ambient conditions

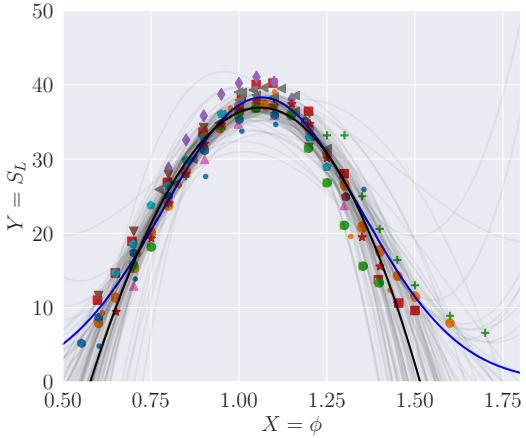
Generalization error: $\mathbf{E}\{\mathcal{E}(f)\} = \underbrace{E(Y, f_B(X))}_{\text{Noise}} + \underbrace{(f_B(X) - \mathbf{E}\{f(X)\})^2}_{\text{Bias}} + \underbrace{\mathbf{E}\left[(\mathbf{E}\{f(X)\} - f(X))^2\right]}_{\text{Variance}}$

1st order regression

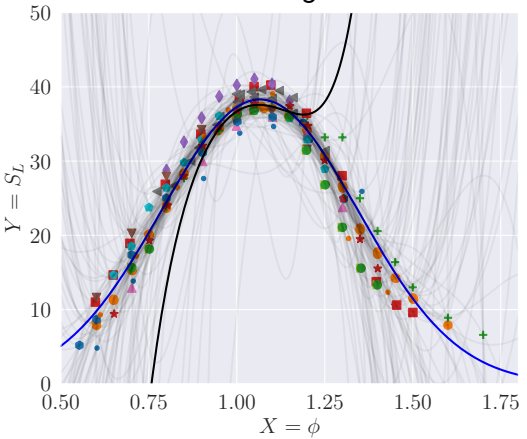


Underfitting: low variance, high bias

3rd order regression



5th order regression



Overfitting: high variance, low bias

Supervised learning methods

- Learning from collection of *labeled* input/output data, $\{(x, y)_i\}_{i=1}^{N}$, to predict outcome: $\hat{y} = f(\hat{x})$
- Learning tasks: regression, classifications
- Algorithms
  - **Logistic regression**
  - **Decision tree and random forest**
  - **Neural networks**
  - Support vector machines

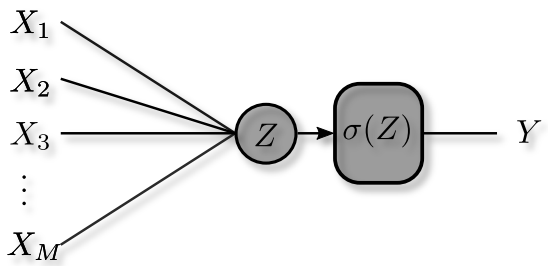
Supervised learning methods

Logistic regression

- Binary classification of learning data $\{(x,y)_i\}_{i=1}^N$ with $y_i \in \{\psi_1 = 0, \psi_2 = 1\}$

$$Y(X) = P(Y = \psi_1 | X) = \sigma(Z)$$

$$Z = w^T X + b$$



with non-linear transfer function

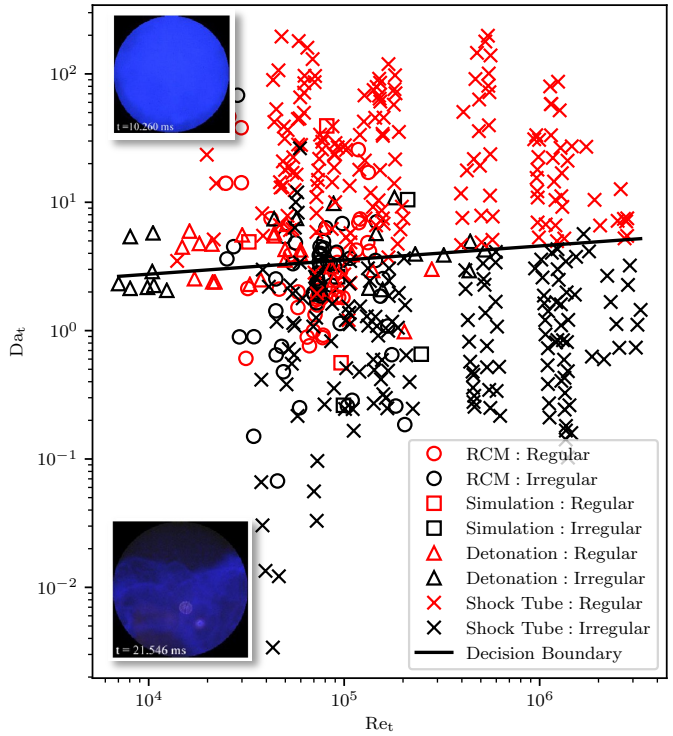
$$\sigma(Z) = \frac{1}{1 + \exp\{-Z\}}$$

- Determine model parameters $\theta = (w, b)^T$ by minimizing likelihood function: $L(y|\theta) = \prod \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i}$

Supervised learning methods

Logistic regression

- Application
  - Classification of irregular combustion ignition conditions

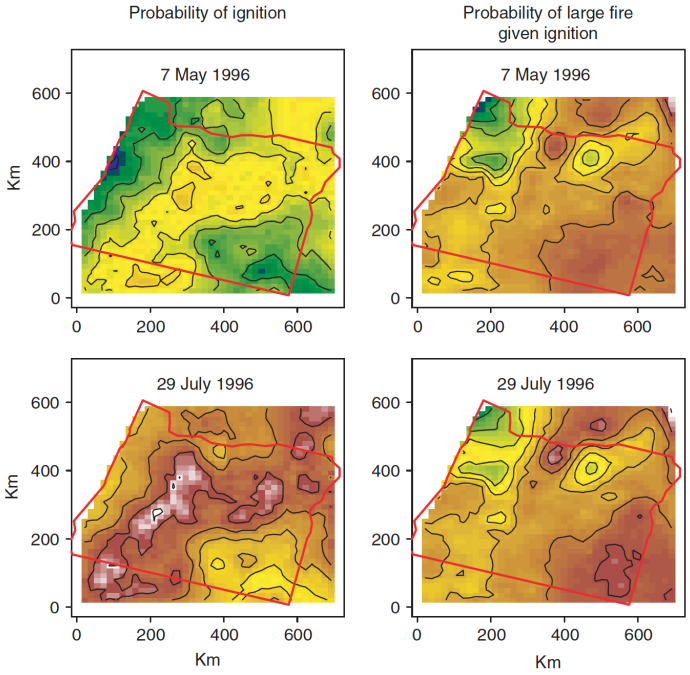


Grogan and Ihme, Shockwave, 28 , 941-954, 2018

Supervised learning methods

Logistic regression

- Application
  - Classification of irregular combustion ignition conditions
  - Prediction of fire occurrence and fire risk assessment

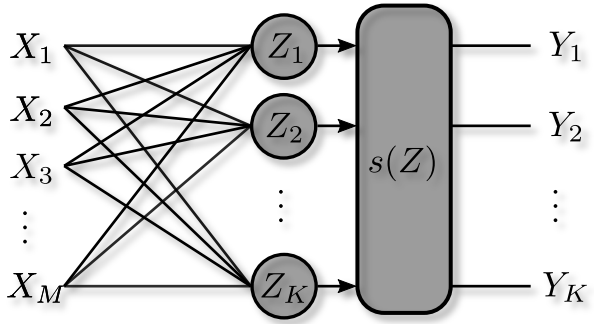


Preisler et al. Int. J. Wildland Fire, 13, 133-142, 2004

Supervised learning methods

Logistic regression

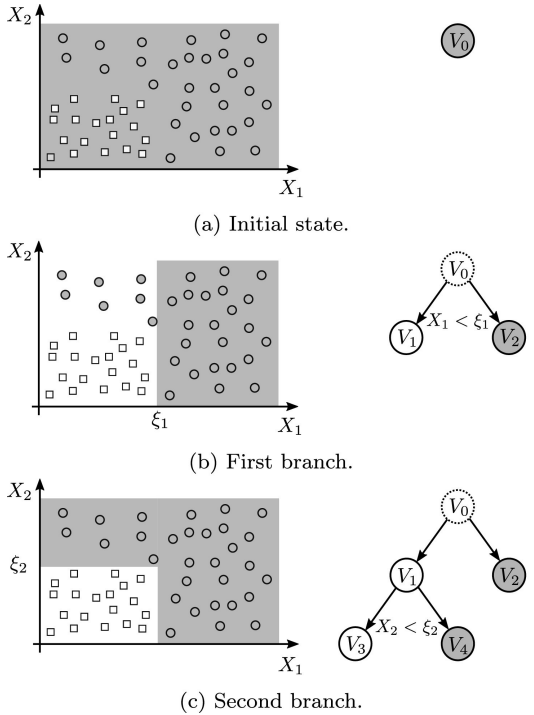
- Remarks and limitations
  - Pre-determined hypothesis
  - Linear separability of feature space
  - Limitation in representing non-linear decision boundaries
  - Fundamental building block of neural networks



Supervised learning methods

Decision trees

- Partition feature space into hypercuboids
- Representation of arbitrarily complex relationships
- Learning and branching based on impurity measure (Gini index)
- High degree of interpretability through feature importance score



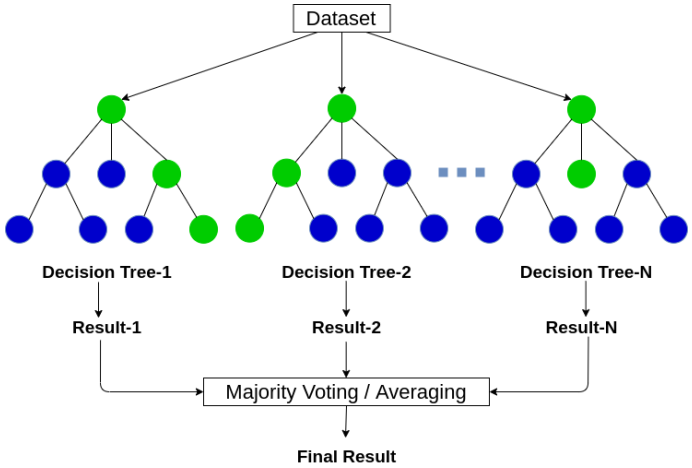
Supervised learning methods

Decision trees

- Challenges: Reduce overfitting

Random Forest

- Strength of weak learnability: Reduce variance and maintain low bias by combining decision trees into random forests
  - Not prone to overfitting
  - Maintain benefits of decision trees
  - Cost scales with number of decision trees!



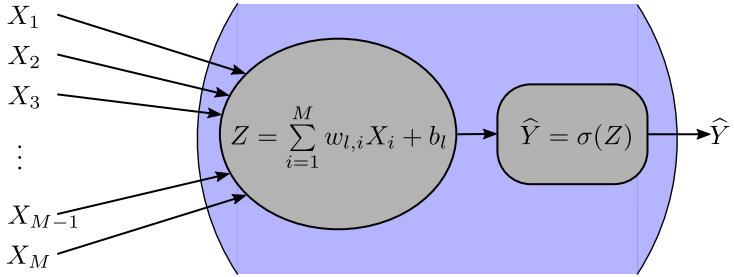
https://ai-pool.com/a/s/random-forests-understanding

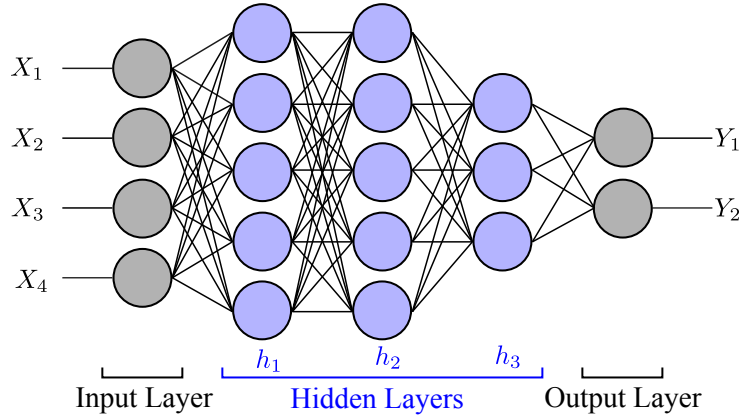
Supervised learning methods

Neural Networks

- Mimics neural brain structure through dense set of inter-connected neurons
- Fundamental building block: logistic regression → Multilayer perceptrons

$$\sigma(Z) = \frac{1}{1 + \exp\{-Z\}}$$

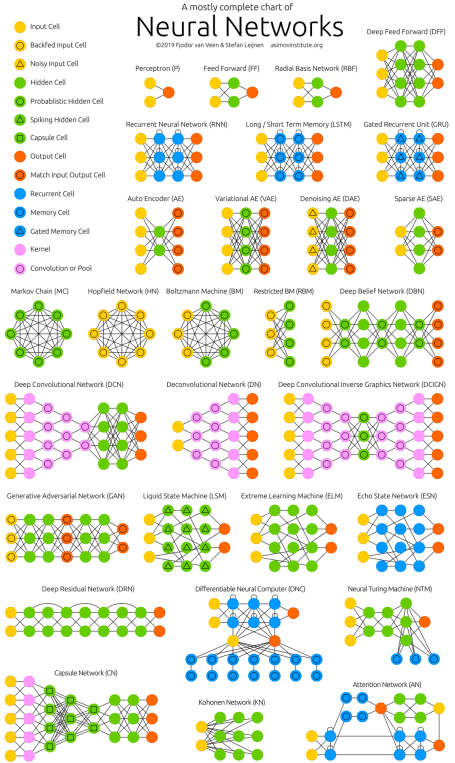




Supervised learning methods

Neural Networks

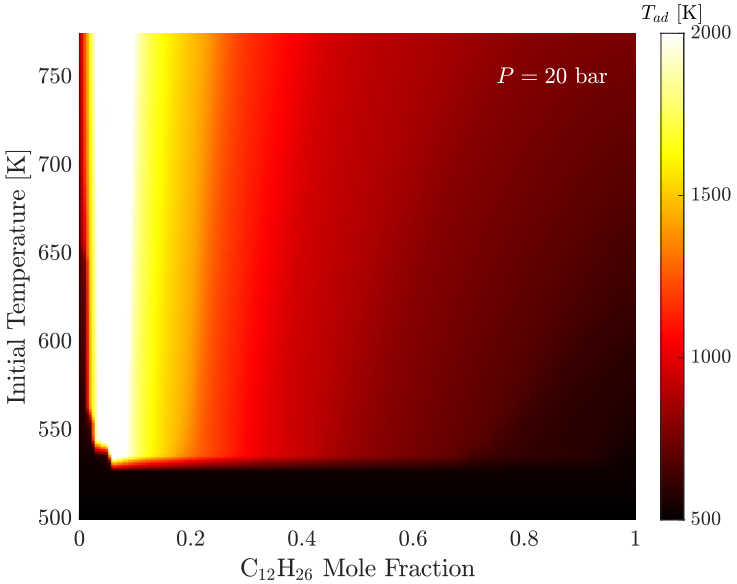
- Specialized network architectures
  - Deep neural networks
  - Convolutional neural networks
  - Recurrent neural networks
  - Graph neural networks

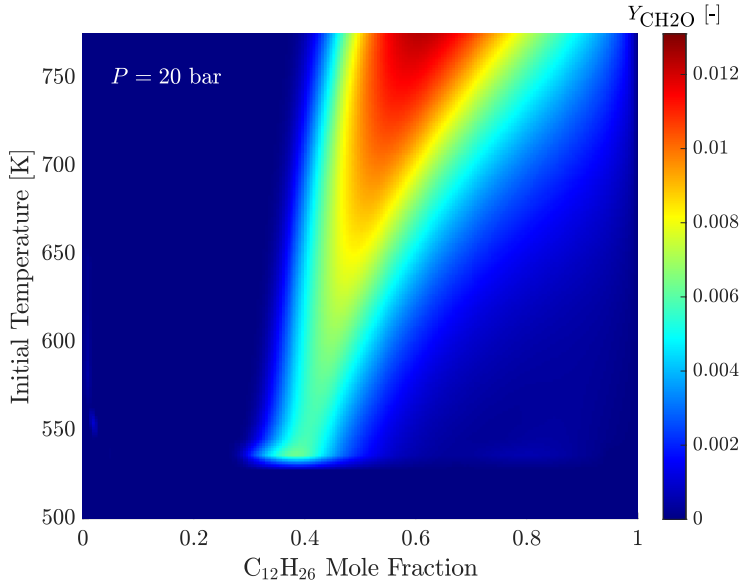


https://www.asimovinstitute.org/neural-network-zoo/

Supervised learning methods

Classification of ignition behavior in $C_{12}H_{26}$/air mixture

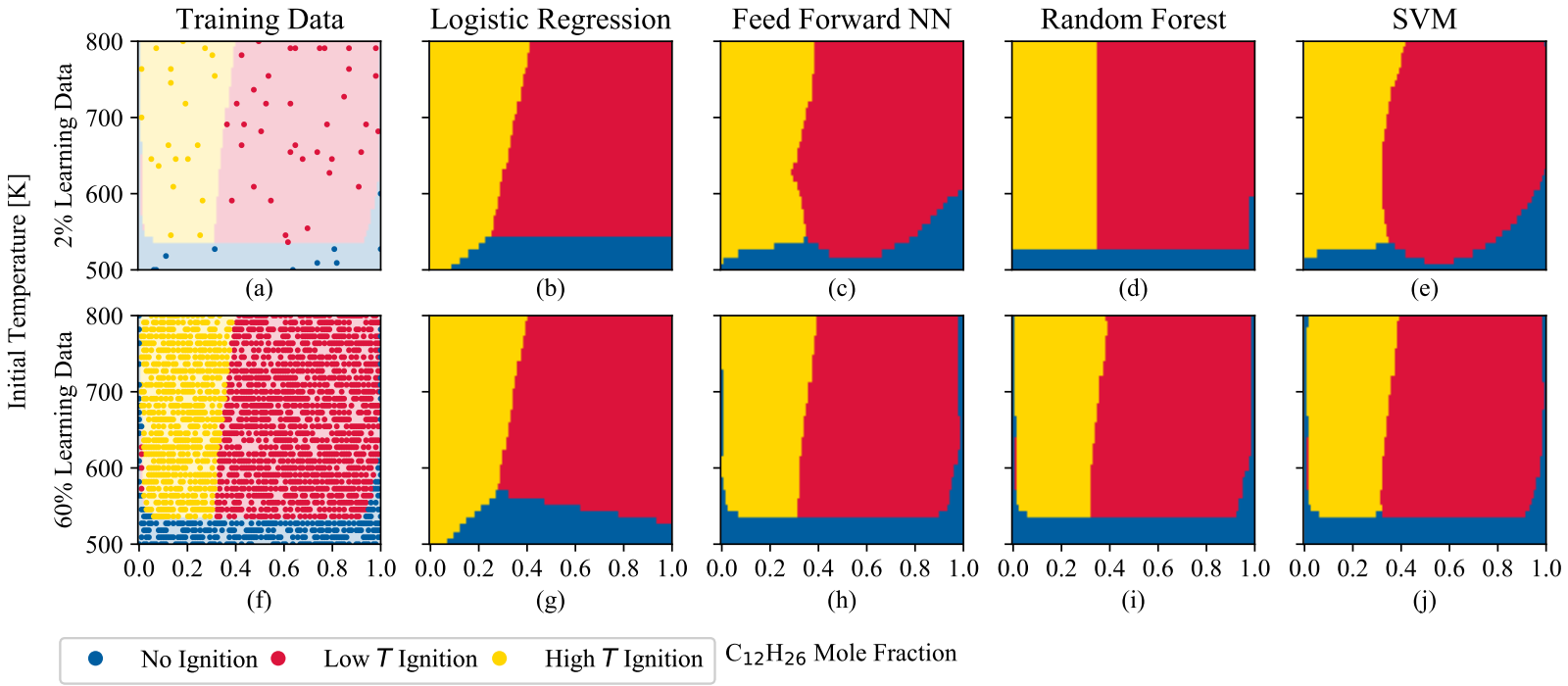




https://github.com/lhmeGroup/CombML_Tutorials

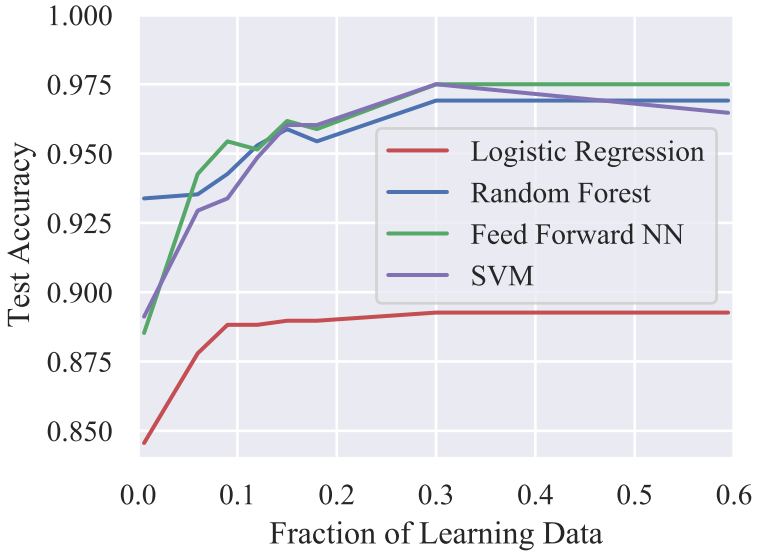
Supervised learning methods

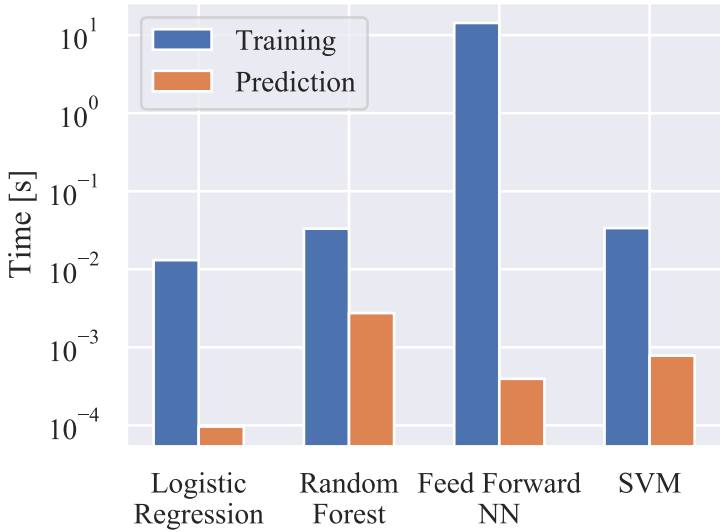
Classification of ignition behavior in $C_{12}H_{26}$/air mixture



Supervised learning methods

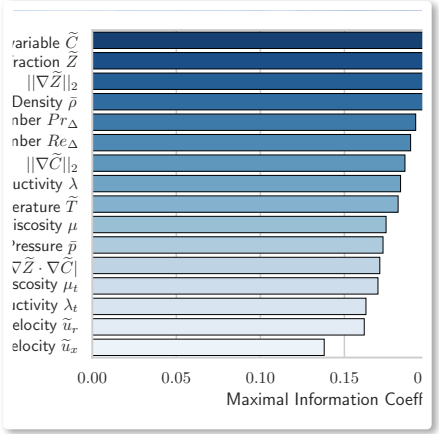
Classification of ignition behavior in $C_{12}H_{26}$/air mixture





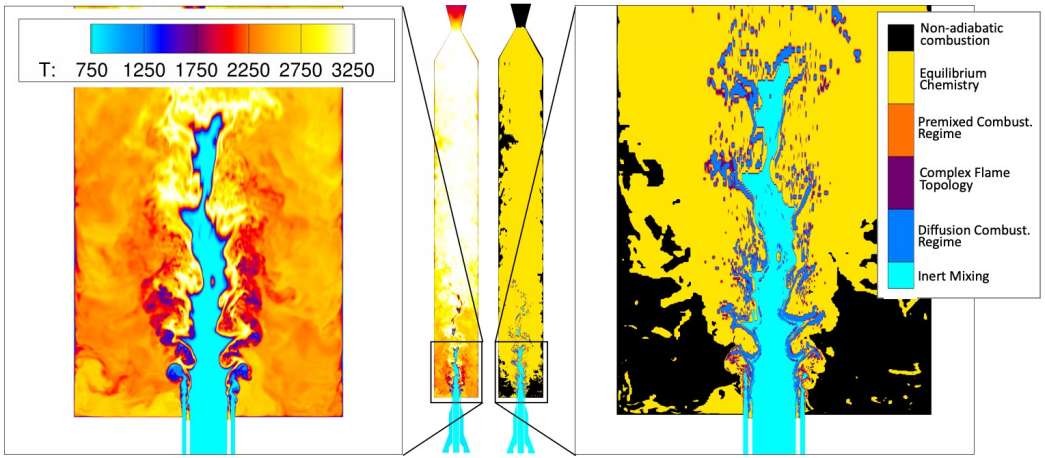
CombML Application

DATA ASSISTED
COMBUSTION MODELING



Motivation

- High-fidelity simulations of turbulent reacting flows can incur high computational costs
- Local assignment of combustion models that are consistent with underlying physics and accurately describe quantities of interest



Motivation

- Large-eddy simulations enable predictions of complex combustion processes through solution of filtered conservation equations:

$$\partial_t \overline{\rho} + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}) = 0$$
$$\partial_t (\overline{\rho}\widetilde{\boldsymbol{u}}) + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{u}}) = -\nabla \cdot (\overline{p}\boldsymbol{I}) + \nabla \cdot (\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}^{sgs})$$
$$\partial_t (\overline{\rho}\widetilde{e}_t) + \nabla \cdot [\widetilde{\boldsymbol{u}}(\overline{\rho}\widetilde{e}_t + \overline{p})] = -\nabla \cdot (\overline{\boldsymbol{q}}_v + \boldsymbol{q}^{sgs}) + \nabla \cdot [(\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}^{sgs}) \cdot \widetilde{\boldsymbol{u}}]$$
$$\partial_t (\overline{\rho}\widetilde{Y}_k) + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}\widetilde{Y}_k) = -\nabla \cdot (\overline{\boldsymbol{j}}_v + \boldsymbol{j}^{sgs}) + \overline{\dot{\omega}}_k \quad \text{where} \quad k = 1, 2, ..., N_s - 1$$

- High computational costs arises from:
  - Many species
  - Multiple scales and chemical stiffness
  - Closure models for turbulence chemistry interaction and turbulent transport

Combustion modeling

Topology-based combustion models

- ✓ Lower computational cost
- ✗ Strong dependency on combustion regime and flame structure
- ✗ Require pre-computation and tabulation
- • Examples: flamelet-type models (FPV, FPI, FGM, etc.)

Topology-free combustion models

- ✓ Weak dependency on combustion regime and flame structure
- ✓ On-the-fly evaluation of modeled species
- ✗ Higher computational cost
- • Examples: DRG, PFA, QSS, PE, RCCE

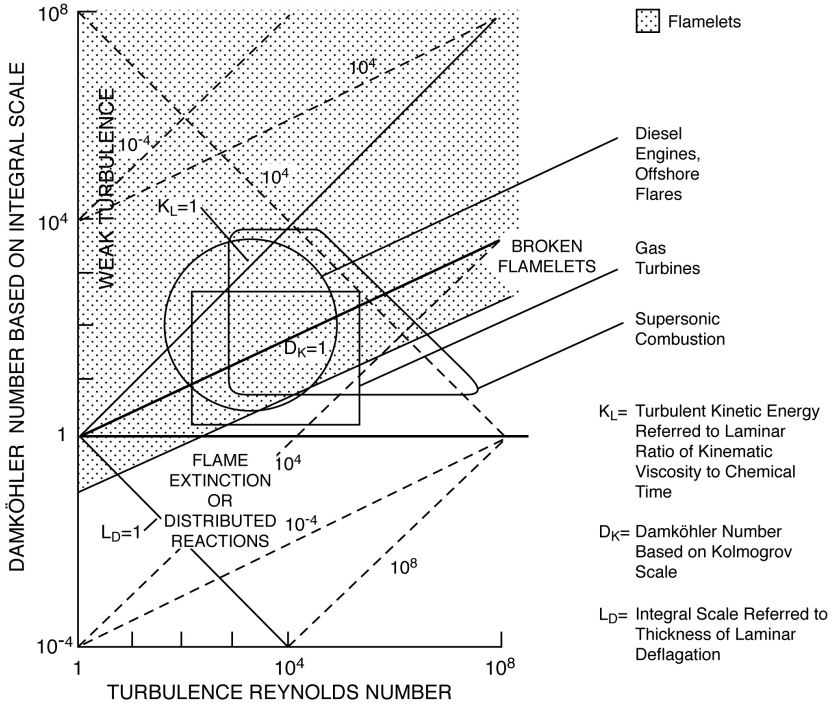


Fig. 2 Regimes of turbulent combustion in a diagram of a turbulence Reynolds number R_ℓ and a Damköhler number D_ℓ, both based on the integral scale of the turbulence.

Williams, AIAA Aero. Sci. Meeting & Exhibit, 2006

Problem with ML-based regression

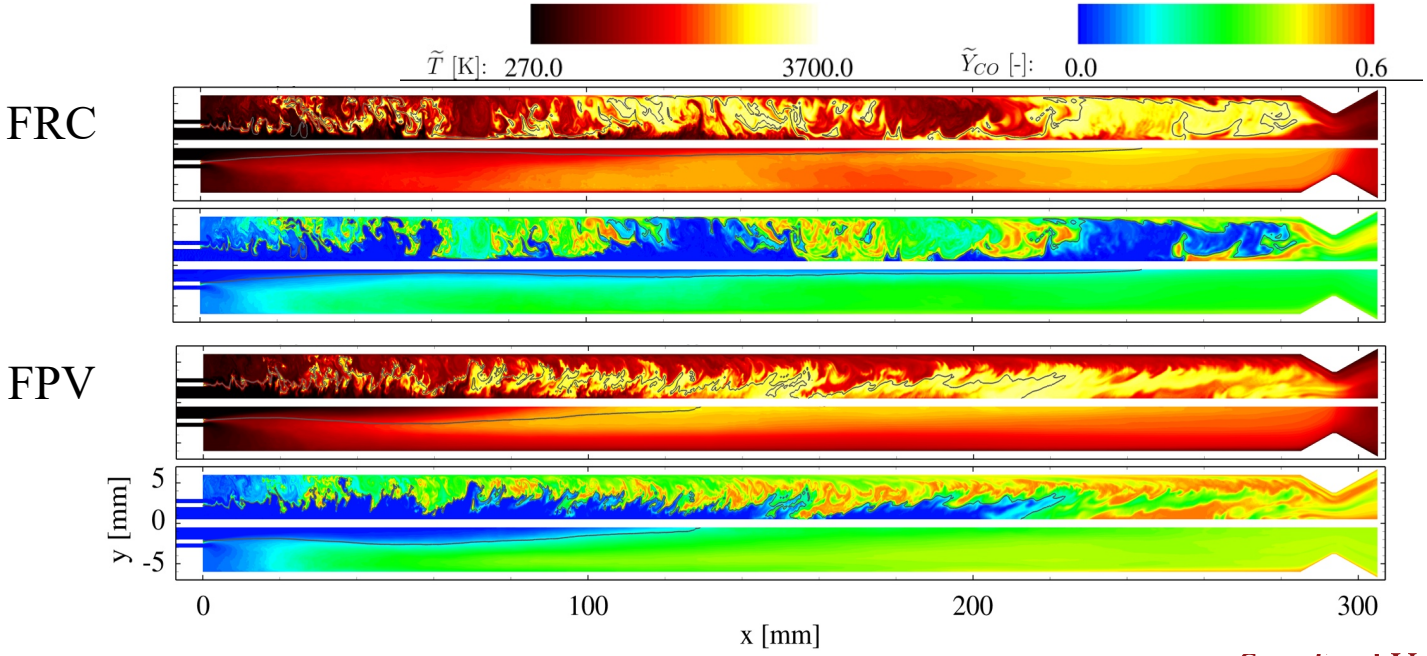
- Physical models versus data-driven models: conservation laws versus complex cross-correlations
- Data-driven models may violate physics during extrapolation tasks
- Data driven models are prone to numerical instability

Solution?

- Use data-driven method to **assist** the selection of low-fidelity physics-based model through classification

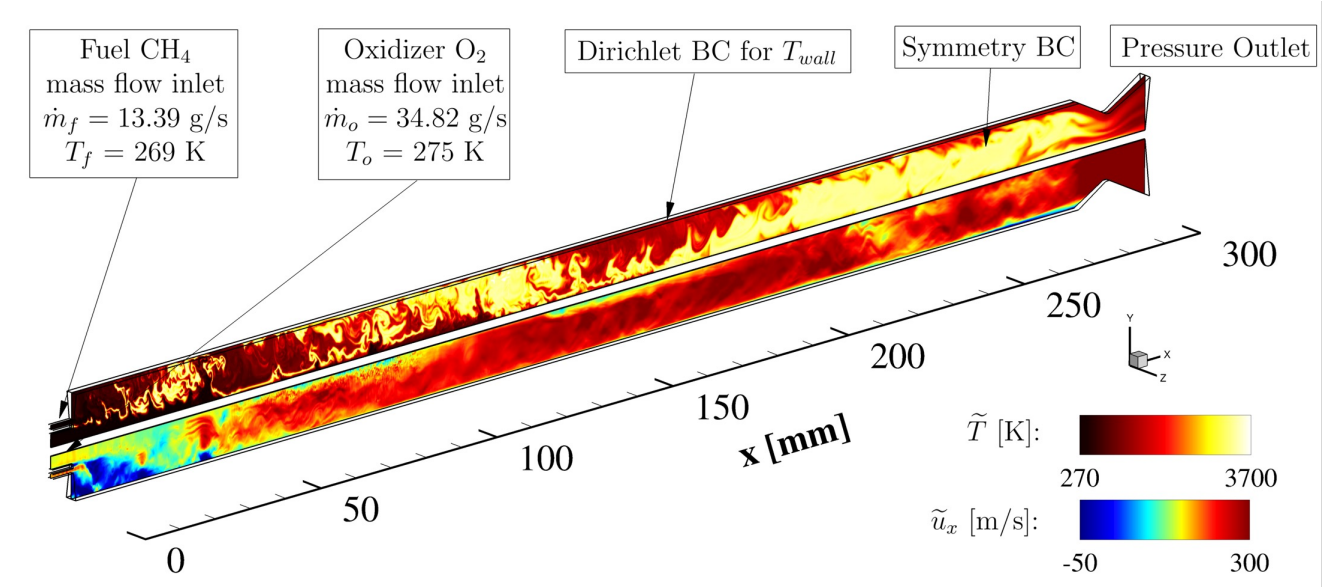
Objective

- Finite rate chemistry (FRC) is very accurate but computationally expensive
- Flamelet Progress Variable (FPV) model cannot capture thermal boundary layers
- Use ML to improve on FPV simulations at a lower cost than FRC.



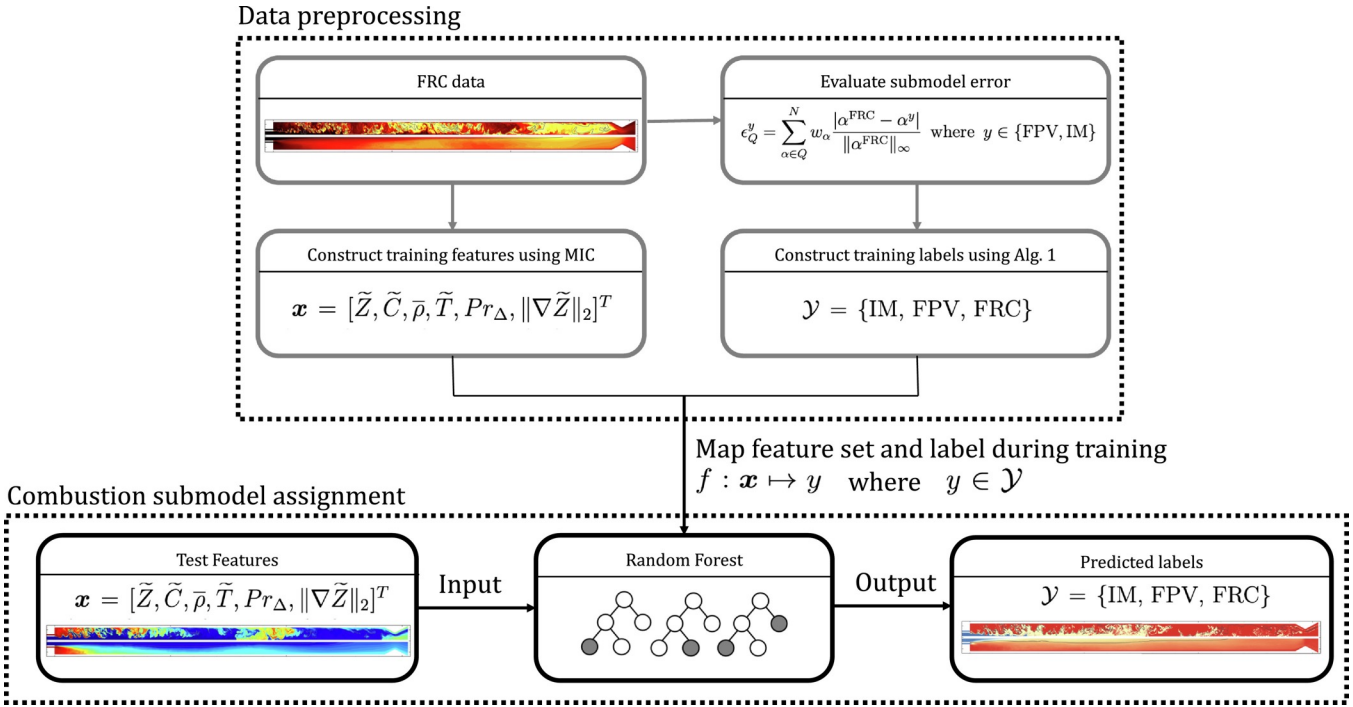
Configuration

- Based on TUM GOX/GCH4 single element rocket combustor
- 2×10^5 cells, axisymmetric domain, with minimum cell size of 30 μm



Silvestri, Celano, Haidn, and Knab, *EUCASS* (2015)

Data-assisted LES



Constructing training labels/output

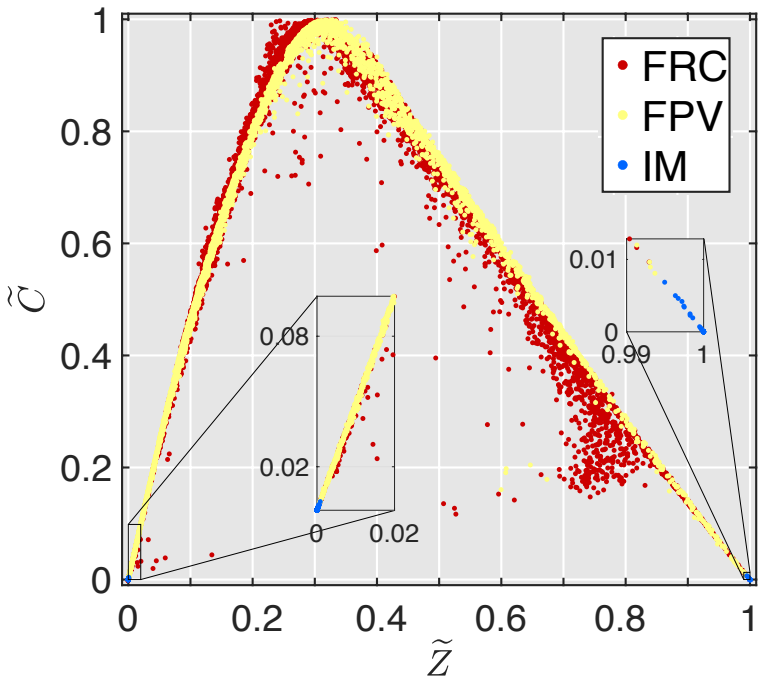
Error between different submodels for $\alpha = \{\tilde{T}, \widetilde{Y_{CO}}\}$

$$\epsilon_Q^y = \sum_{\alpha \in Q} w_\alpha \frac{|\alpha^{\mathrm{FRC}} - \alpha^y|}{\|\alpha^{\mathrm{FRC}}\|_\infty} \quad \text{with} \quad y \in \{\mathrm{FPV}, \mathrm{IM}\}$$

User defined threshold

if $\epsilon_Q^{IM} < \theta_Q^{IM}$ **then**
  use inert mixing (IM)
else if $\epsilon_Q^{FPV} < \theta_Q^{FPV}$ **then**
  use tabulated chemistry (FPV)
else
  use finite-rate chemistry (FRC)
end

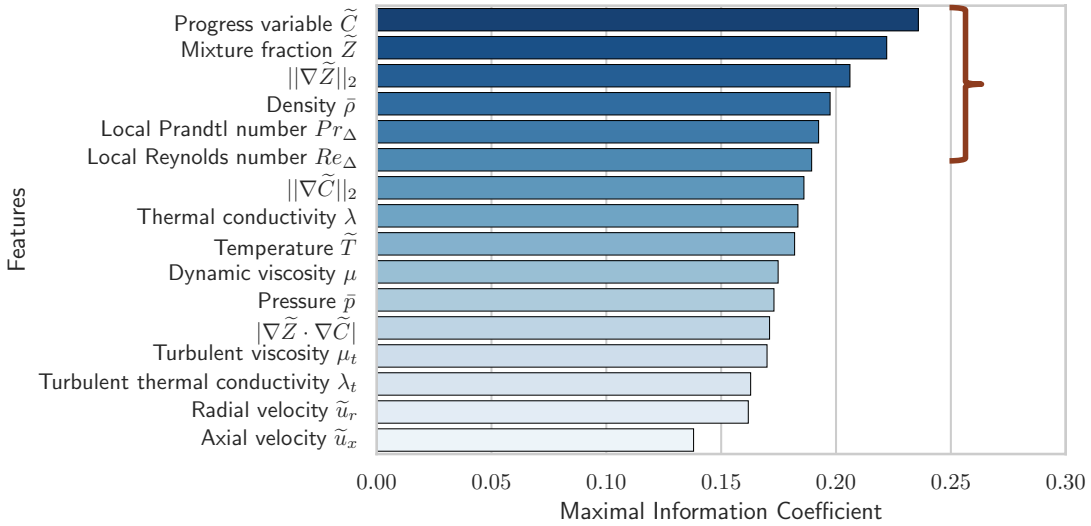
Training data for $\theta_{\{T,\mathrm{CO}\}} = 0.02$



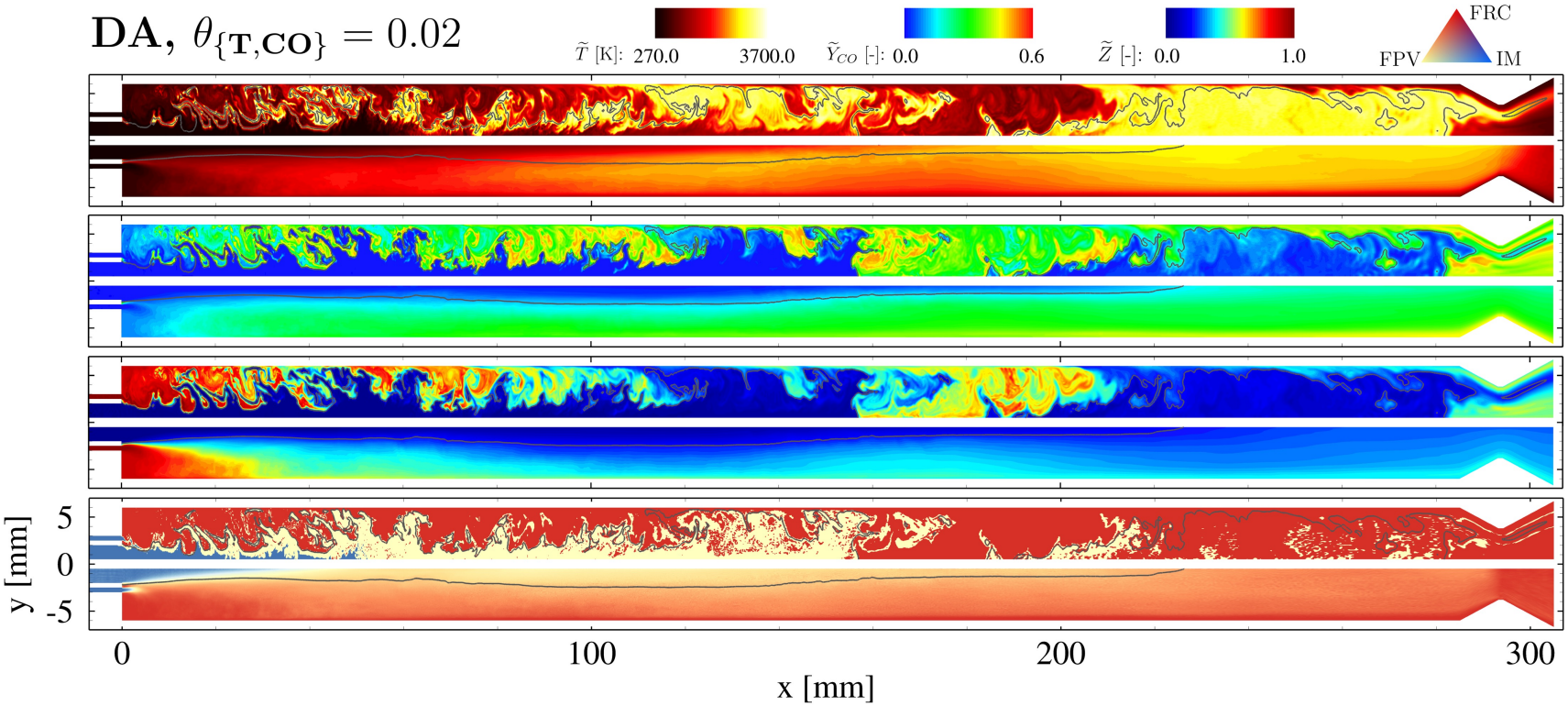
Feature/input selection

- Use Maximal Information Coefficient (MIC) to select most relevant input
- MIC is a correlation measure for nonlinear data, similar to R2 measure

MIC relating features with $\epsilon^{FPV}_{\{T,CO\}}$

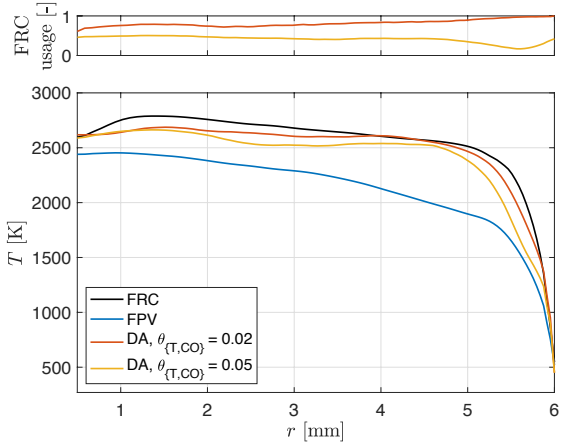


A posteriori results

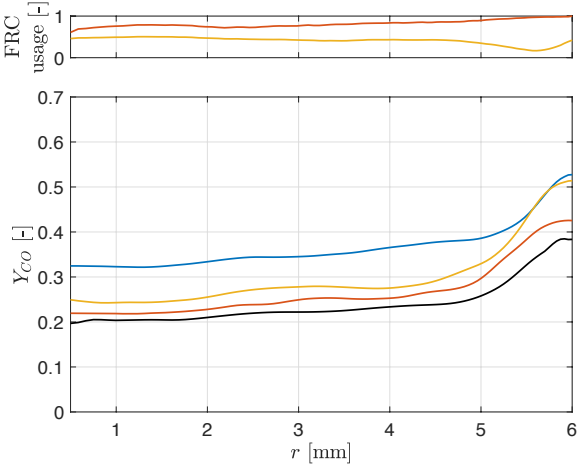


A posteriori results

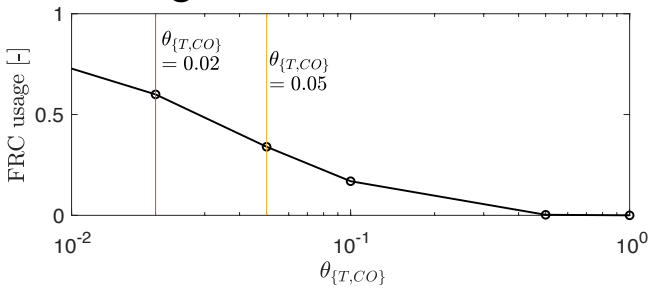
Temperature vs radius (x = 280 mm)



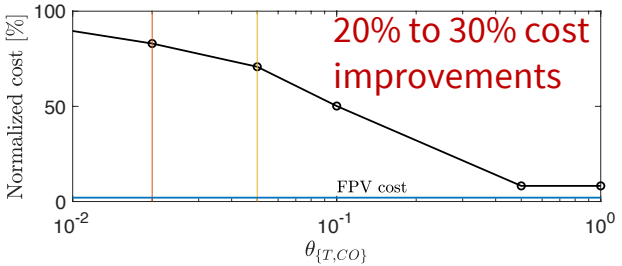
CO mass fraction vs radius (x = 280 mm)



FRC usage vs user-defined threshold

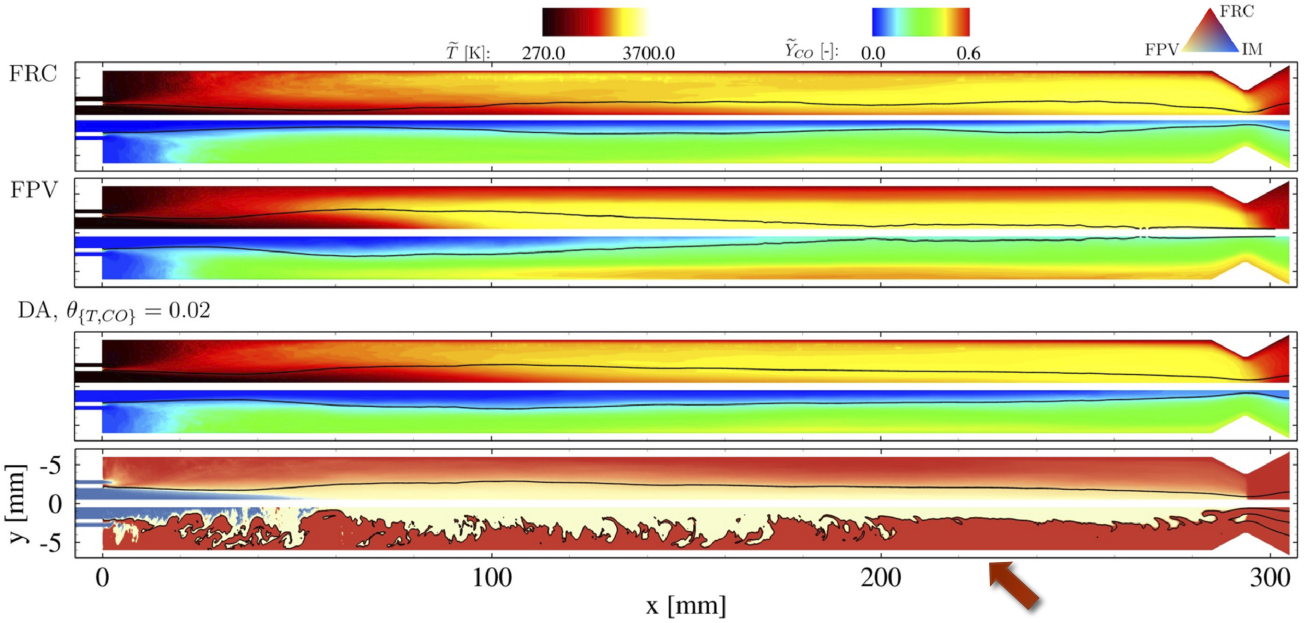


Cost vs user-defined threshold



A posteriori results: Extrapolation task?

- For a modified configuration with 3x inlet mass flow rate
- Method demonstrates ability to generalize for different configurations



Data-assisted combustion model

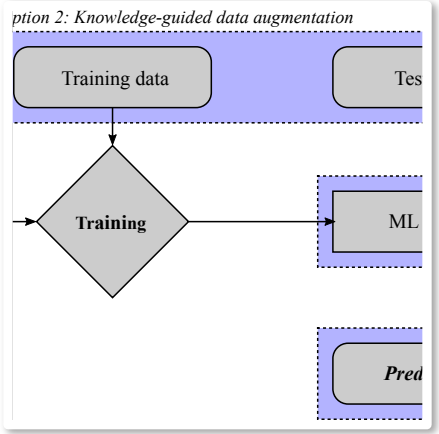
- Classification can be used to locally assign combustion models
- Overall desired fidelity can be controlled during labelling
- DA simulations outperform monolithic FPV and monolithic FRC simulations in accuracy and cost, respectively
- Random forest and neural networks demonstrate similar accuracy *a priori*, but show different behavior
- Neural networks have a more contiguous assignment than random forests *a posteriori.*

Further advances

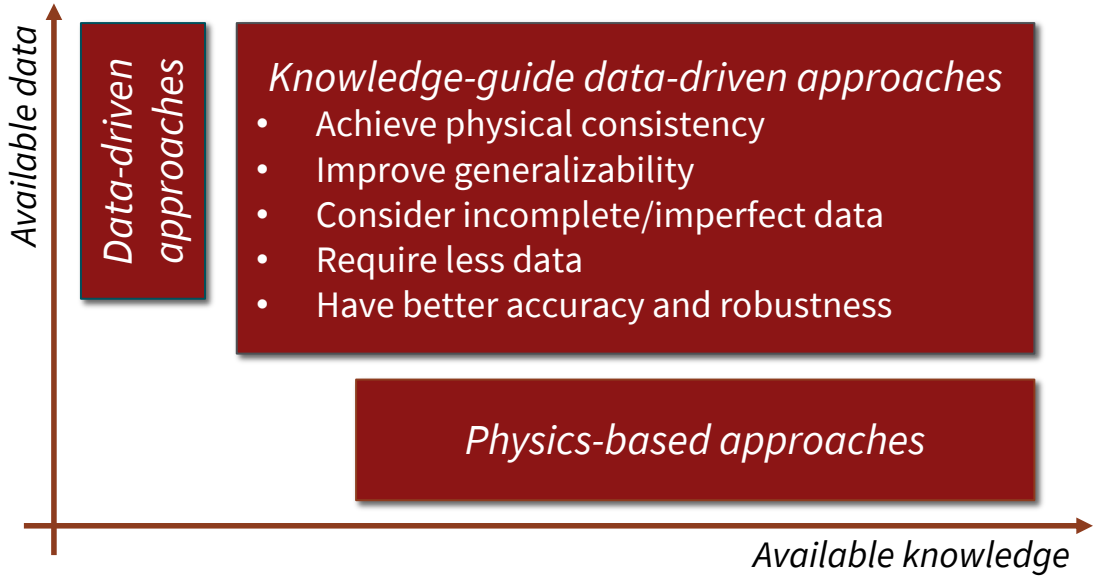
- Inclusion of generic physical laws during labeling and training
- Extend submodel selection to turbulence modeling, and equation of state.

Modifying ML for Combustion Science and Engineering

KNOWLEDGE-GUIDED DATA-DRIVEN METHODS

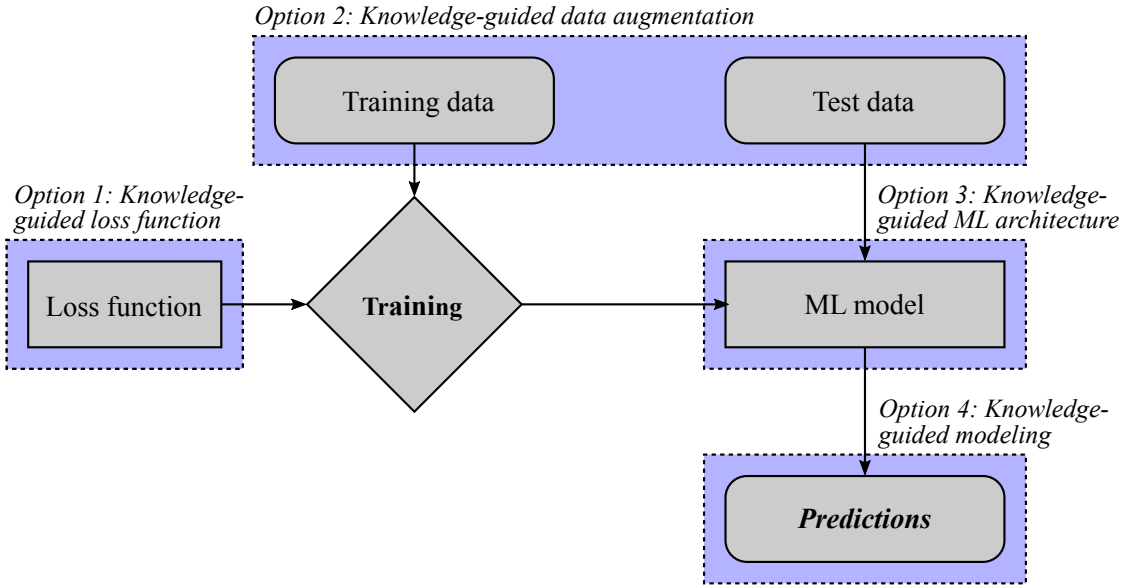


Knowledge-guided ML for combustion science and engineering



Knowledge-guided ML for combustion science and engineering

Leverage domain knowledge to reduce dependence on (label) data



Motivation

- Large-eddy simulations enable predictions of complex combustion processes through solution of filtered conservation equations:

$$\partial_t \overline{\rho} + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}) = 0$$
$$\partial_t (\overline{\rho}\widetilde{\boldsymbol{u}}) + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{u}}) = -\nabla \cdot (\overline{p}\boldsymbol{I}) + \nabla \cdot (\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}^{sgs})$$
$$\partial_t (\overline{\rho}\widetilde{e}_t) + \nabla \cdot [\widetilde{\boldsymbol{u}}(\overline{\rho}\widetilde{e}_t + \overline{p})] = -\nabla \cdot (\overline{\boldsymbol{q}}_v + \boldsymbol{q}^{sgs}) + \nabla \cdot [(\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}^{sgs}) \cdot \widetilde{\boldsymbol{u}}]$$
$$\partial_t (\overline{\rho}\widetilde{Y}_k) + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}\widetilde{Y}_k) = -\nabla \cdot (\overline{\boldsymbol{j}}_v + \boldsymbol{j}^{sgs}) + \overline{\dot{\omega}}_k \quad \text{where} \quad k = 1, 2, ..., N_s - 1$$

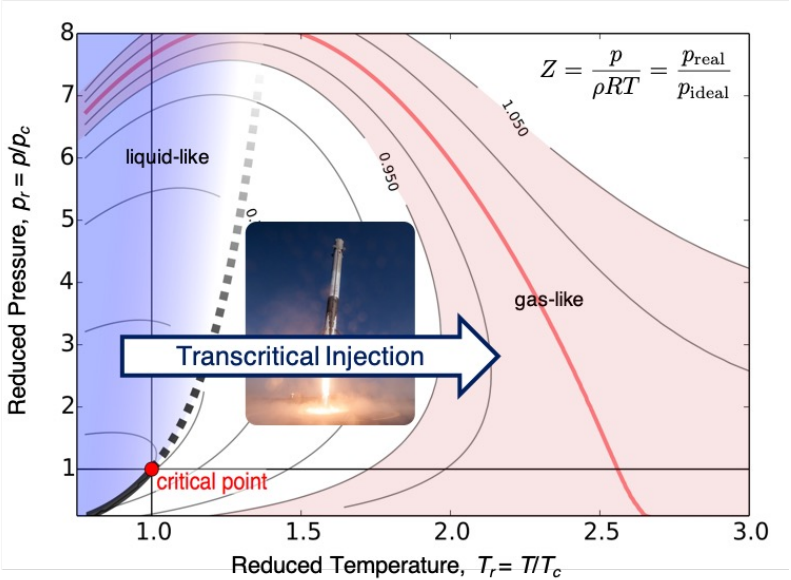
- High computational costs arises from:
  - Many species
  - Multiple scales and chemical stiffness
  - Closure models for turbulence chemistry interaction and turbulent transport

Motivation

Develop SGS closure models for high-pressure combustion environments using ML methods?

Can we extract physical insight from ML methods?

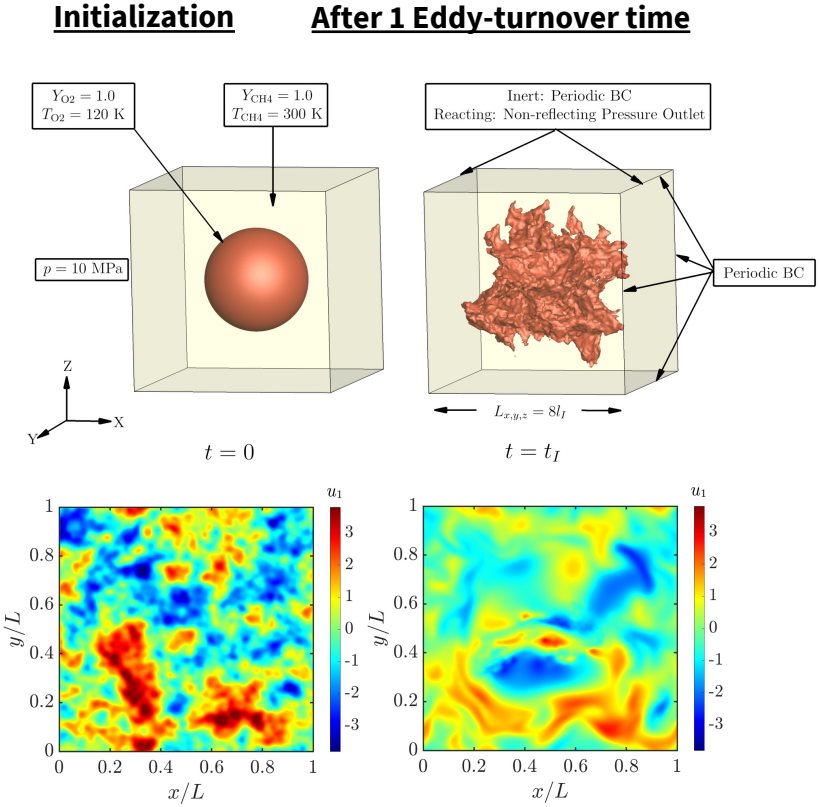
- LOX-GCH4 engines operate under transcritical condition
- SGS Models traditionally been developed for incompressible flows.



Problem formulation

Computational Setup

- DNS: 128^3 domain and $\Delta x = O(\eta_K)$
- Diffuse Interface Method
- Peng-Robinson Cubic EoS
- CH4/O2 chemistry: 5-species
- Initial conditions: von Karman/Pao spectrum and decaying turbulence
- Initial scalar flowfield initialized with 1D counterflow diffusion flame



[1] P. C. Ma, Y. Lv, M. Ihme, *J. Comput. Phys*. (2017)
[2] D.-Y.Peng,D.B.Robinson, *Ind.Eng.Chem.Fundam*. (1976)
[3] B. Franzelli, E. Riber, L. Y. Gicquel, T. Poinsot, *Combust. Flame* (2012)
[4] T.Saad,D.Cline,R.Stoll,J.C.Sutherland, *AIAA J.* (2017)

A priori analysis: Steps

- Apply box-filtering on DNS:

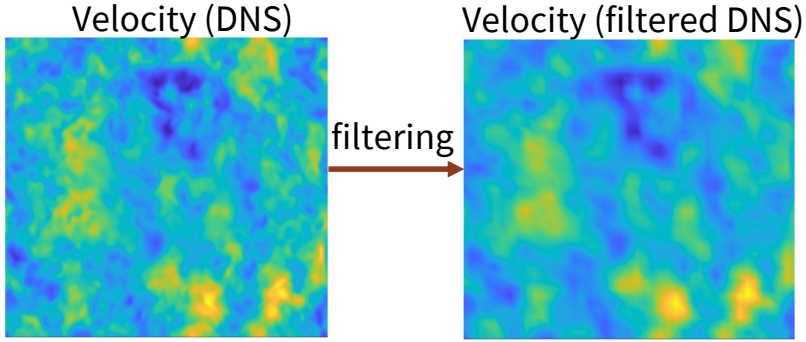
$$\overline{\phi(\boldsymbol{x})} = \int_V \phi(\boldsymbol{x}) H(\boldsymbol{x}-\boldsymbol{y}, \overline{\Delta}) d\boldsymbol{y}$$

$$H(\boldsymbol{x}-\boldsymbol{y}, \overline{\Delta}) = \begin{cases} \frac{1}{\overline{\Delta}} & \text{for } |\boldsymbol{x}-\boldsymbol{y}| \leq \overline{\Delta}/2, \\ 0 & \text{otherwise.} \end{cases}$$

- Extract exact term:

$$\tau_{ij}^{sgs} = \overline{\rho}(\widetilde{u_i u_j} - \widetilde{u}_i \widetilde{u}_j)$$

- Model SGS term with ML models
  - Trained at $t = \tau$
  - Tested at $t = 1.5\tau$



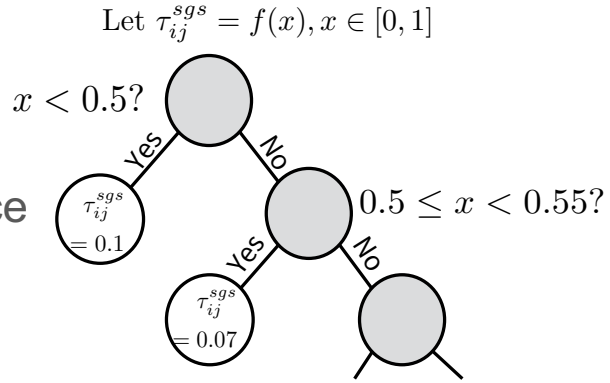
Input (Features)	Output (Labels)
$\widetilde{u}_i, \frac{\partial \widetilde{u}_i}{\partial x_j}, \frac{\partial^2 \widetilde{u}_i}{\partial x_j x_k}$	τ_{ij}^{sgs}

- Compare modeled SGS terms with exact terms with Pearson Correlation
  - 0 = Weak Correlation, 1 = Strong correlation

Random forests: Ensemble of Decision Trees

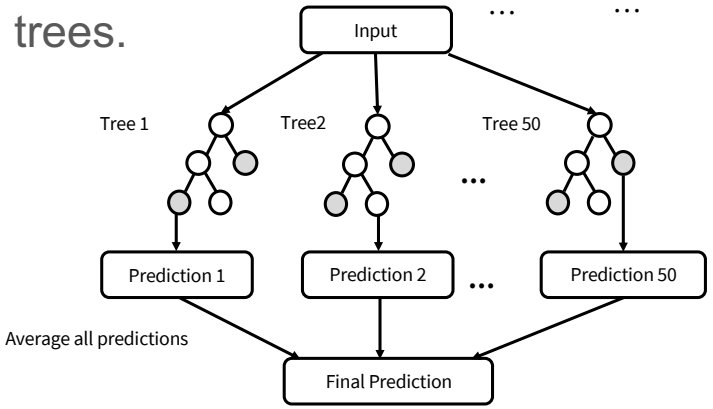
Decision Trees

- Consider a case where subgrid-scale stress can be modelled with a single parameter/feature
- Decision Trees are trained by splitting feature space
- Consider 30 features

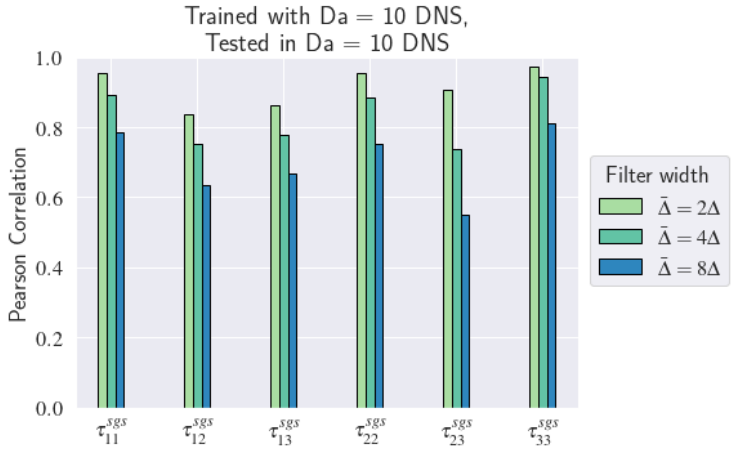


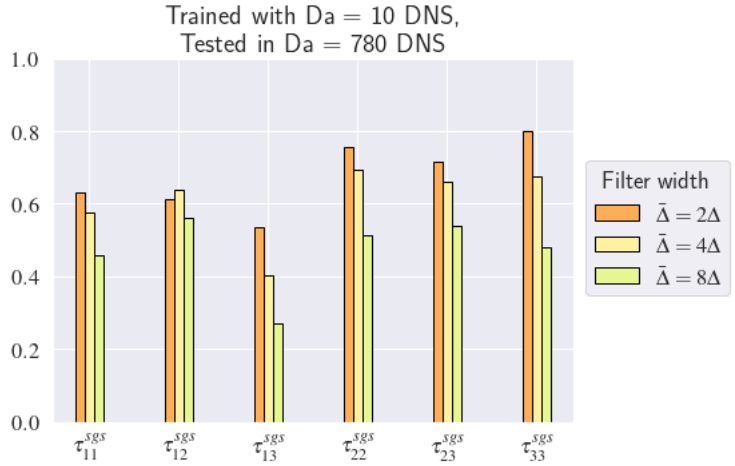
Random Forests

- Random forests are ensembles of decision trees.
- Ensemble-averaging reduces overfitting.
- Cost-effective regression method.
  - Training cost: 88s using 1 CPU
  - Inference cost: 2.4s using 1 CPU
- Can provide feature importance



Random forest Model Predictions





- Tested Random forest SGS stresses a priori over three filter widths.
- Random forest SGS shows excellent performance when trained and tested on own configuration
- Results do not perform well in out-of-distribution configuration.
- Random forests provide a **feature importance score** through a mean decrease in node impurity (variance).

Sparse symbolic regression for model discovery

Represent SGS model as an $\mathbf{Ax} = \mathbf{b}$ or $\mathbf{X}\boldsymbol{\beta} = \mathbf{f}$ problem.

$\widetilde{u}_i, \dfrac{\partial \widetilde{u}_i}{\partial x_j}, \dfrac{\partial^2 \widetilde{u}_i}{\partial x_j x_k}$ $\quad [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_k]^T$ $\quad \tau_{ij}^{sgs}$

Introduce nonlinearities:

$$G^d(\mathbf{X})\boldsymbol{\beta} = \mathbf{f}$$

$$\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1n} & X_{11}^2 & X_{11}X_{12} & \cdots & X_{1n}^d \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{m1} & X_{mn} & \cdots & X_{mn} & X_{m1}^2 & X_{m1}X_{m2} & \cdots & X_{mn}^d \end{bmatrix}$$

Solve it as an optimization problem:

$$\min_{\beta} \left\{ \frac{1}{m} ||\mathbf{f} - G^d(\mathbf{X})\boldsymbol{\beta}||_2^2 || \right\}$$

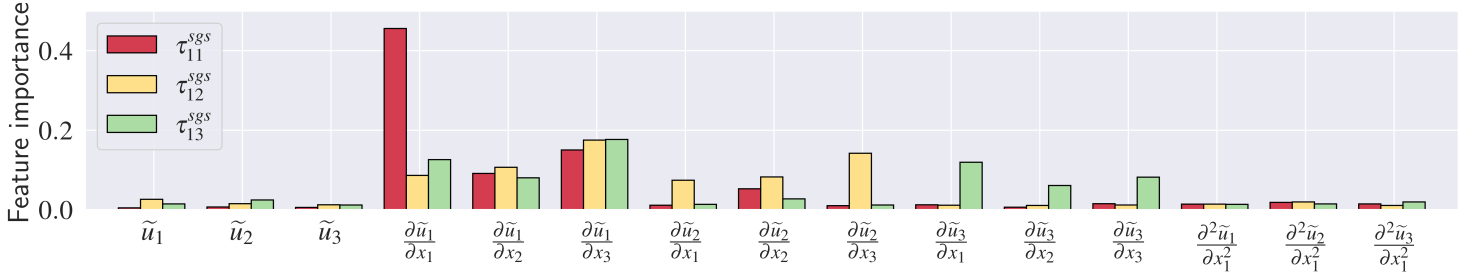
Add L1 norm constraint for sparsity:

$$\min_{\beta} \left\{ \frac{1}{m} ||\mathbf{f} - G^d(\mathbf{X})\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}_1|| \right\}$$

Find $\boldsymbol{\beta}$ via any optimization scheme gradient descent.

Sparse regression for model discovery

- Can be computationally expensive as cost scales with MN^d
- Feature importance is measured through mean decrease in node impurity



Feature importance simplifies sparse symbolic regression for model discovery

$$\frac{\tau_{ij}^{sgs}}{\bar{\rho}u'^2} = f_{ij}\left[G^{d=2}\left(\frac{\widetilde{u}_i}{u'}, \frac{\bar{\Delta}}{u'}\frac{\partial \widetilde{u}_i}{\partial x_j}, \frac{\bar{\Delta}}{u'}\frac{\partial \widetilde{u}_j}{\partial x_i}, \frac{\bar{\Delta}^2}{u'}\frac{\partial^2 \widetilde{u}_i}{\partial x_j \partial x_k}, \frac{\bar{\Delta}^2}{u'}\frac{\partial^2 \widetilde{u}_j}{\partial x_i \partial x_k}, \frac{\bar{\Delta}^2}{u'}\frac{\partial^2 \widetilde{u}_k}{\partial x_i \partial x_i}\right)\right]$$

$$\frac{\tau_{ij}^{sgs}}{\bar{\rho}u'^2} = f_{ij}\left[G^{d=2}\left(\frac{\bar{\Delta}}{u'}\frac{\partial \widetilde{u}_i}{\partial x_k}, \frac{\bar{\Delta}}{u'}\frac{\partial \widetilde{u}_j}{\partial x_k}\right)\right]$$

Sparse regression for model discovery

$$\frac{\tau_{ij}^{sgs}}{\bar{\rho}u'^2} = f_{ij}\left[G^{d=2}\left(\frac{\bar{\Delta}}{u'}\frac{\partial \widetilde{u}_i}{\partial x_k}, \frac{\bar{\Delta}}{u'}\frac{\partial \widetilde{u}_j}{\partial x_k}\right)\right]$$

Apply sparse symbolic regression

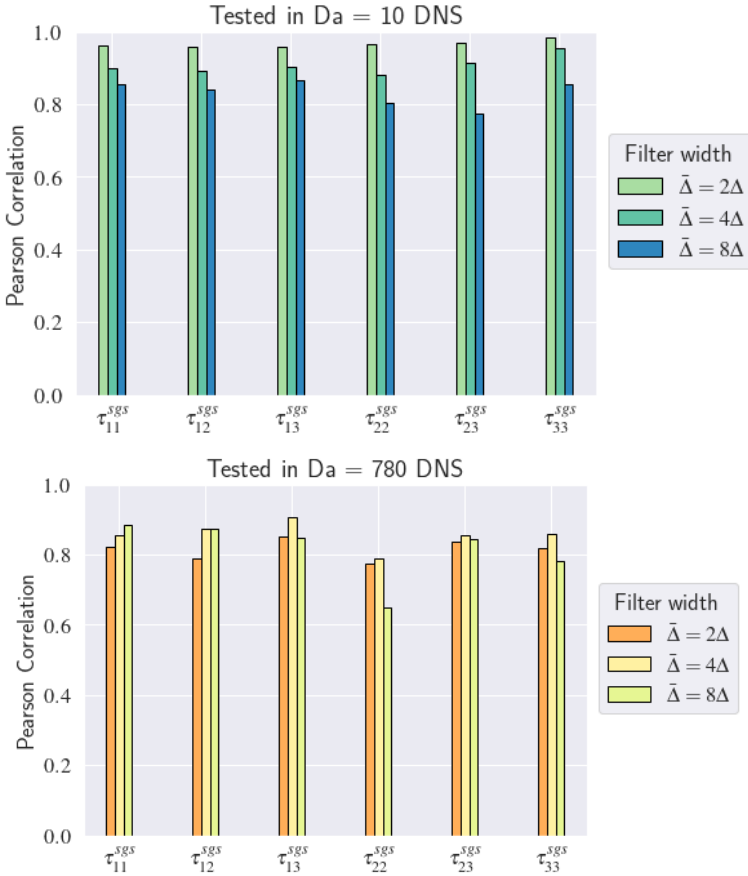
$$\tau_{11}^{sgs} \simeq \bar{\rho}\bar{\Delta}^2\left(0.116\frac{\partial \widetilde{u}_1}{\partial x_1}\frac{\partial \widetilde{u}_1}{\partial x_1} + 0.191\frac{\partial \widetilde{u}_1}{\partial x_2}\frac{\partial \widetilde{u}_1}{\partial x_2} + 0.207\frac{\partial \widetilde{u}_1}{\partial x_3}\frac{\partial \widetilde{u}_1}{\partial x_3}\right)$$

$$\vdots$$

$$\tau_{33}^{sgs} \simeq \bar{\rho}\bar{\Delta}^2\left(0.251\frac{\partial \widetilde{u}_3}{\partial x_1}\frac{\partial \widetilde{u}_3}{\partial x_1} + 0.177\frac{\partial \widetilde{u}_3}{\partial x_2}\frac{\partial \widetilde{u}_3}{\partial x_2} + 0.124\frac{\partial \widetilde{u}_3}{\partial x_3}\frac{\partial \widetilde{u}_3}{\partial x_3}\right)$$

$$\tau_{ij}^{sgs} \simeq \underbrace{\bar{\rho}C_x}_{\sim 1/6}\bar{\Delta}^2\frac{\partial \widetilde{u}_i}{\partial x_k}\frac{\partial \widetilde{u}_j}{\partial x_k}$$

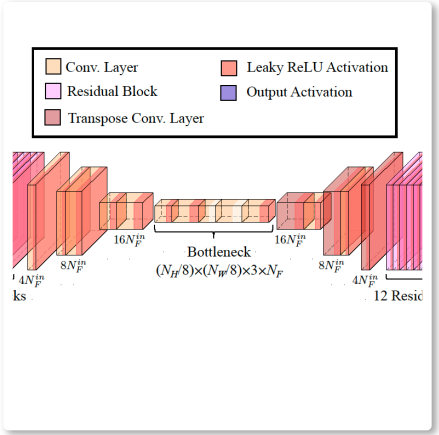
Rediscovered Clark and Ferziger's Gradient Model!



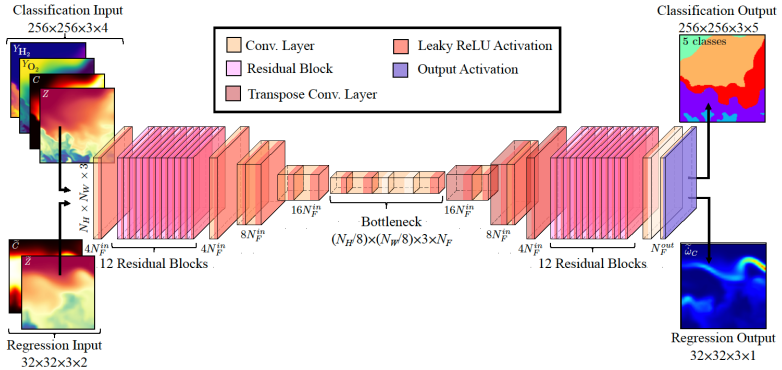
Interpretable ML for SGS model construction

- Demonstrate ability of extracting physically meaningful insight from data
- Random forest models SGS stresses well when trained in-sample
- Random forest does not model SGS stresses well when trained out-of-sample
- Random forest are interpretable and can reveal most important inputs
- Random forest feature importance can be used with spare symbolic regression for mode model discovery

Research Opportunities and Needs



Research Opportunities and needs



- Benchmark datasets and metrics
- Interpretability and explainability
- Quantifying uncertainties of CombML models
- Evaluation out-of-distribution predictions
- Integrating domain knowledge in CombML
- Computational complexity and accuracy

What makes a good dataset?

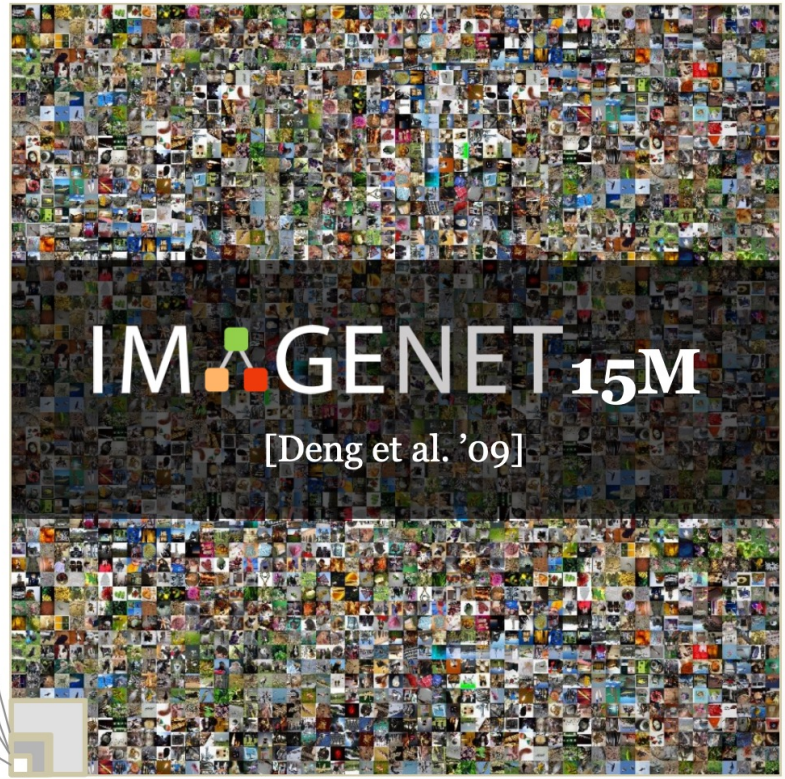
SCALE!!

SUN, 131K
[Xiao et al. '10]

LabelMe, 37K
[Russell et al. '07]

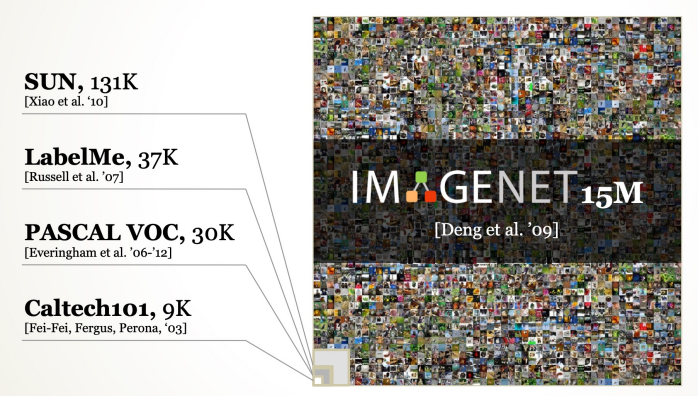
PASCAL VOC, 30K
[Everingham et al. '06-'12]

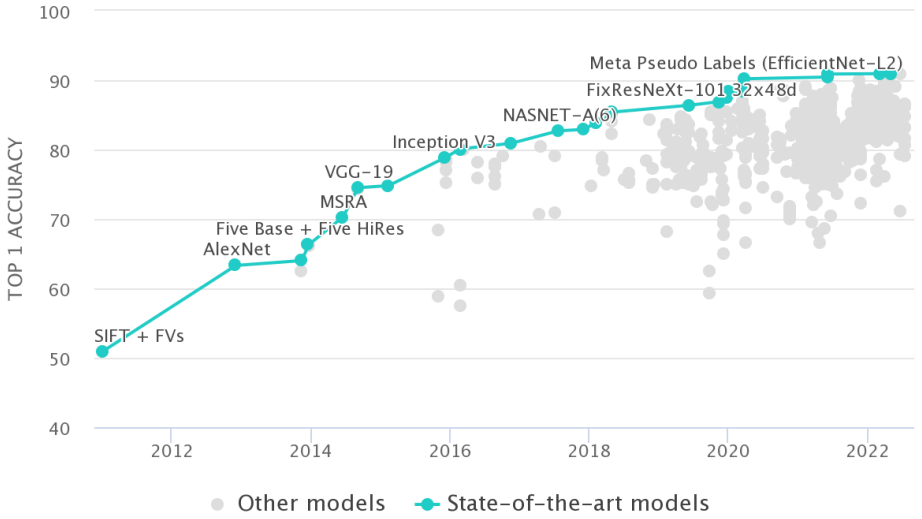
Caltech101, 9K
[Fei-Fei, Fergus, Perona, '03]



What makes a good dataset?

Need for community effort

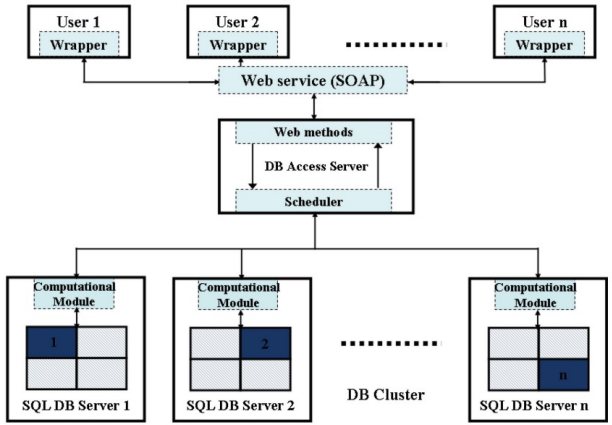




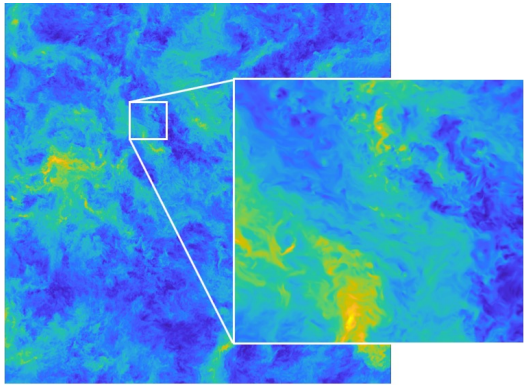
State-of-the-art Flow Physics Scientific Dataset

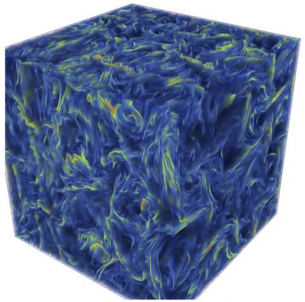
- Johns Hopkins Turbulence Database
- Only 9 3D configurations
- Total Size = 500 TB (3D, Detailed)
- State of the art snapshots > 100 GB

Hosted on dedicated infrastructure



Extremely Detailed!



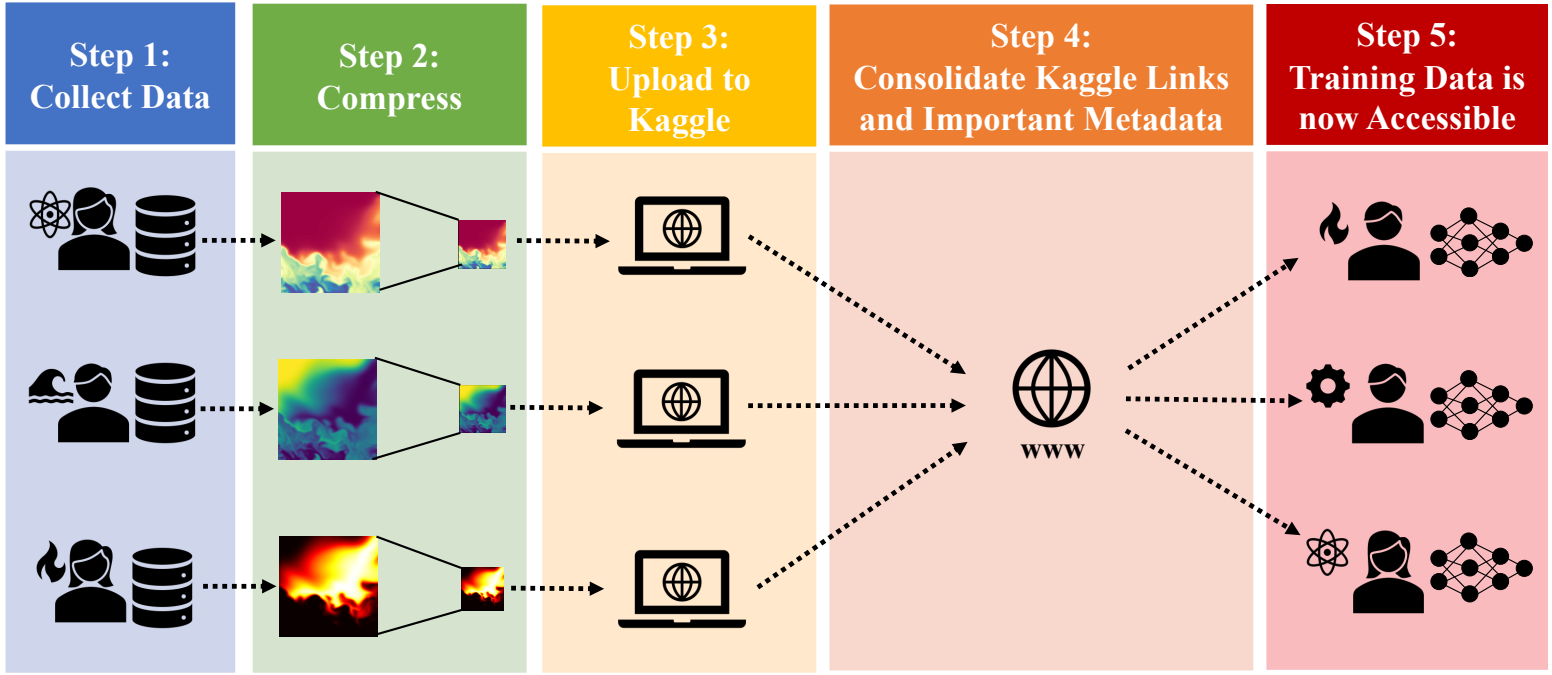


Current field $\nabla \times \mathbf{B}$

http://turbulence.pha.jhu.edu/

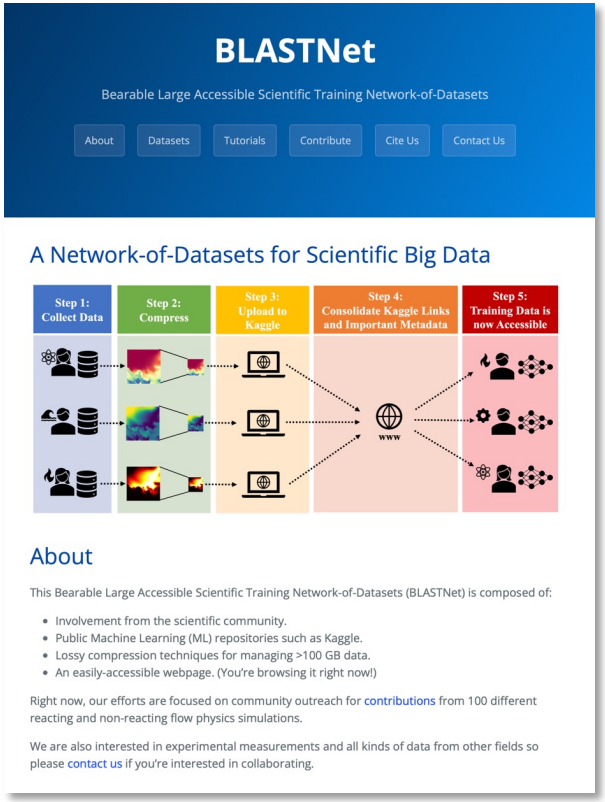
BLASTNet: A community web-platform for big data

- Scale through community involvement/outreach.



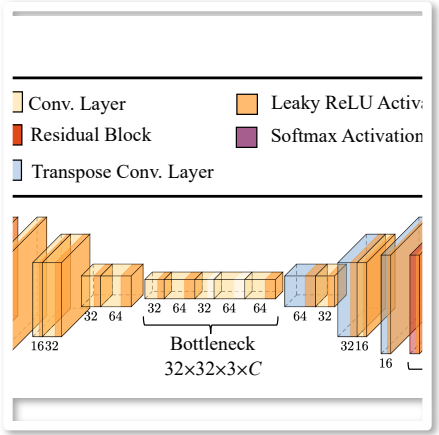
BLASTNet features

- First iteration of BLASTNet aims to **curate 100 different non-reacting flow configurations**
- **Provides tutorials** for sharing and accessing data, ML-samples
- **Stores metadata** in consistent JSON format
- **Provides standards and guidelines** for shared data
- **Hosts discussion forum** for user support and community feedback
- Growing list-of-authors to include each contributor for **fair attribution**



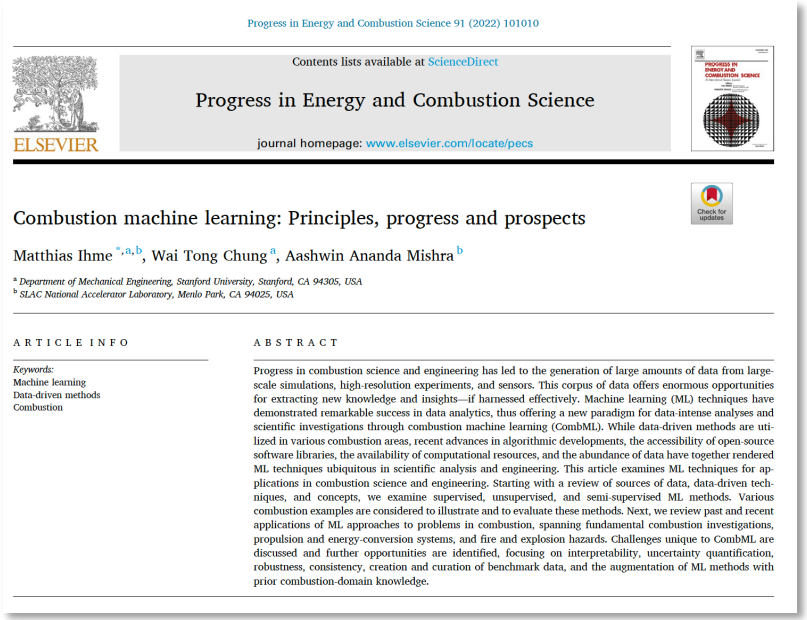
https://blastnet.github.io

Summary


Conv. Layer
Residual Block
Transpose Conv. Layer
Leaky ReLU Activ
Softmax Activation
32 64
32 64 32 64 64
64 32
1632
3216
16
Bottleneck
32×32×3×C

Summary

- Substantial progress in embracing ML for traditional applications
  - Regression, dimensional reduction
  - Control, optimization, modeling
- Adaptation of ML-techniques holds exciting promise for advancing current state-of-the-art in combustion science and engineering
- CombML complements traditional physics-driven inquiry
- Address and embrace challenges unique to combustion science and engineering
  - Uncertainty quantify, interpretability
  - Safety
  - Physical principles



Progress in Energy and Combustion Science 91 (2022) 101010

Contents lists available at ScienceDirect

Progress in Energy and Combustion Science

journal homepage: www.elsevier.com/locate/pecs

Combustion machine learning: Principles, progress and prospects

Matthias Ihme [*,a,b], Wai Tong Chung [a], Aashwin Ananda Mishra [b]

[a] Department of Mechanical Engineering, Stanford University, Stanford, CA 94305, USA
[b] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

ARTICLE INFO

Keywords:
Machine learning
Data-driven methods
Combustion

ABSTRACT

Progress in combustion science and engineering has led to the generation of large amounts of data from large-scale simulations, high-resolution experiments, and sensors. This corpus of data offers enormous opportunities for extracting new knowledge and insights—if harnessed effectively. Machine learning (ML) techniques have demonstrated remarkable success in data analytics, thus offering a new paradigm for data-intense analyses and scientific investigations through combustion machine learning (CombML). While data-driven methods are utilized in various combustion areas, recent advances in algorithmic developments, the accessibility of open-source software libraries, the availability of computational resources, and the abundance of data have together rendered ML techniques ubiquitous in scientific analysis and engineering. This article examines ML techniques for applications in combustion science and engineering. Starting with a review of sources of data, data-driven techniques, and concepts, we examine supervised, unsupervised, and semi-supervised ML methods. Various combustion examples are considered to illustrate and to evaluate these methods. Next, we review past and recent applications of ML approaches to problems in combustion, spanning fundamental combustion investigations, propulsion and energy-conversion systems, and fire and explosion hazards. Challenges unique to CombML are discussed and further opportunities are identified, focusing on interpretability, uncertainty quantification, robustness, consistency, creation and curation of benchmark data, and the augmentation of ML methods with prior combustion-domain knowledge.

https://github.com/IhmeGroup/CombML_Tutorials


Thank you!