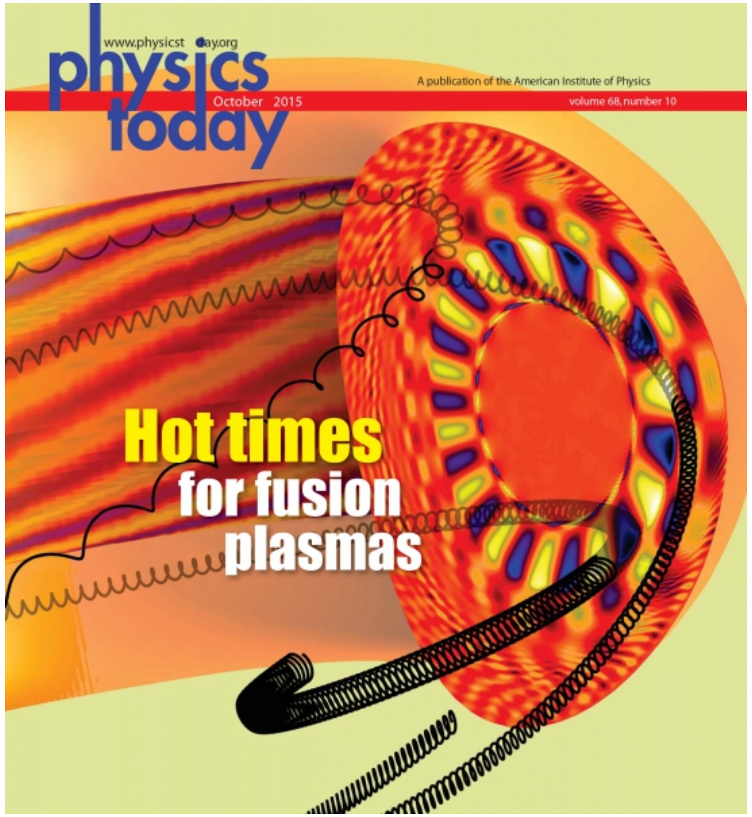
Combining Computing and Learning for (Plasma) Physics

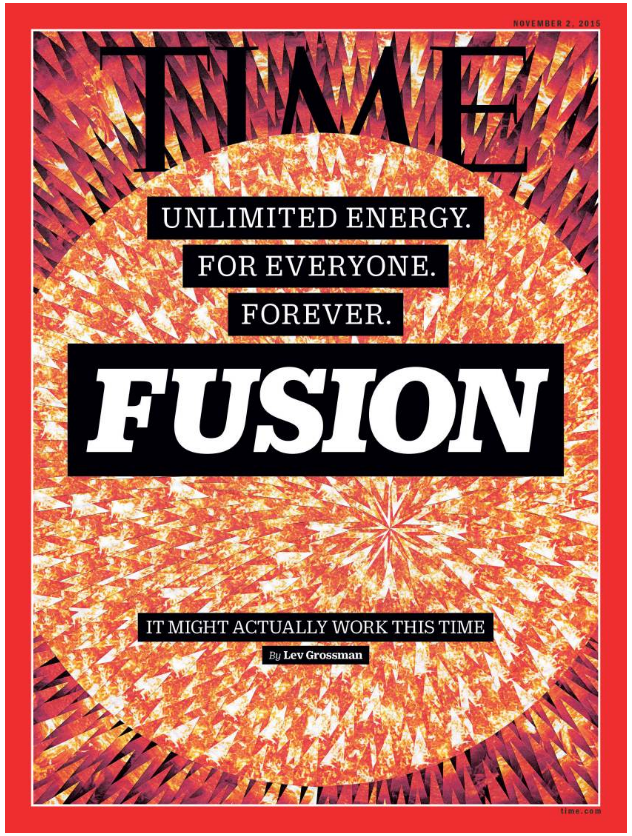
Frank Jenko

Max Planck Institute for Plasma Physics, Garching
Technical University of Munich
The University of Texas at Austin

Workshop on *HPC and Data Science for Scientific Discovery*
IPAM, University of California, Los Angeles, USA, October 15-19, 2018

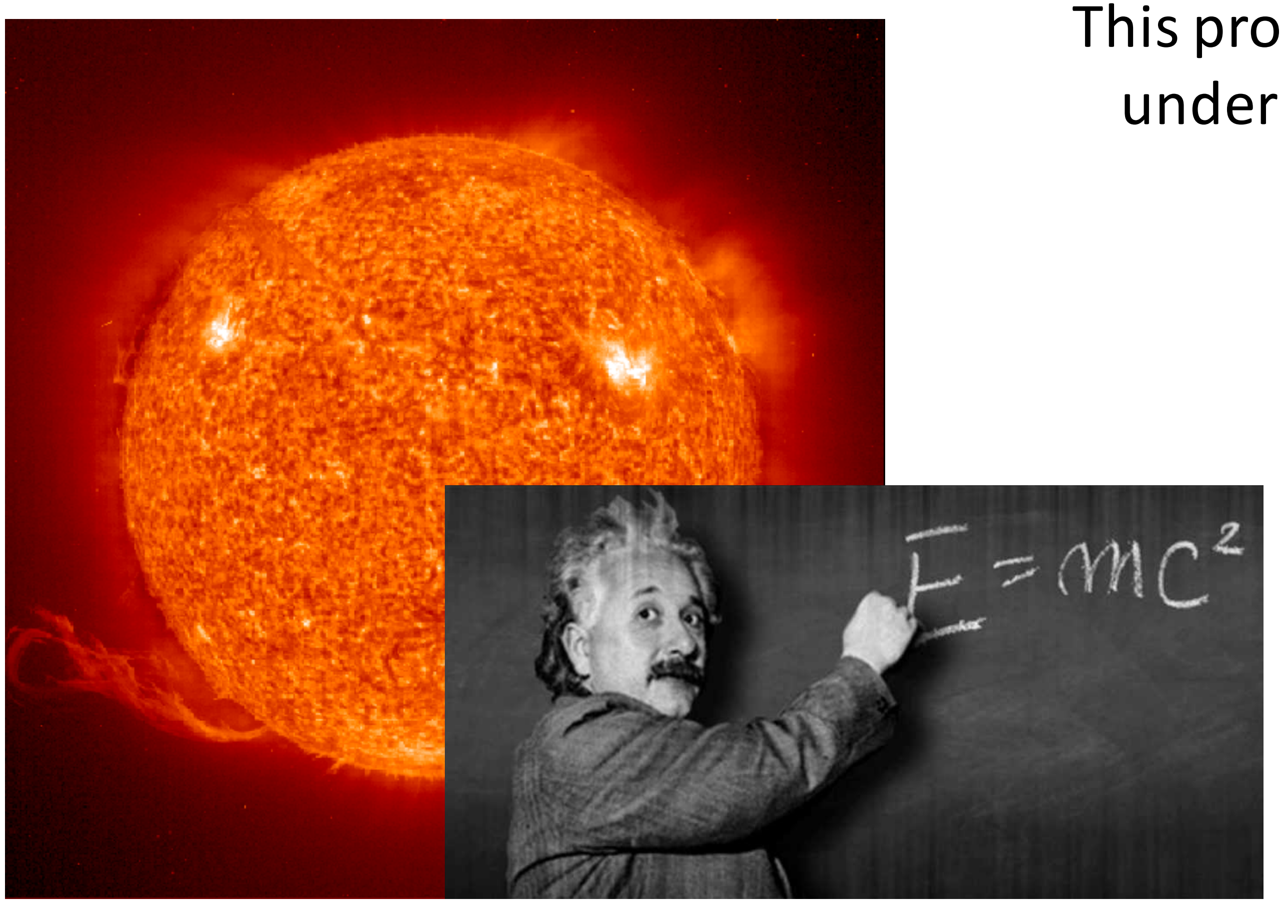
Fusion energy in the news (two examples)



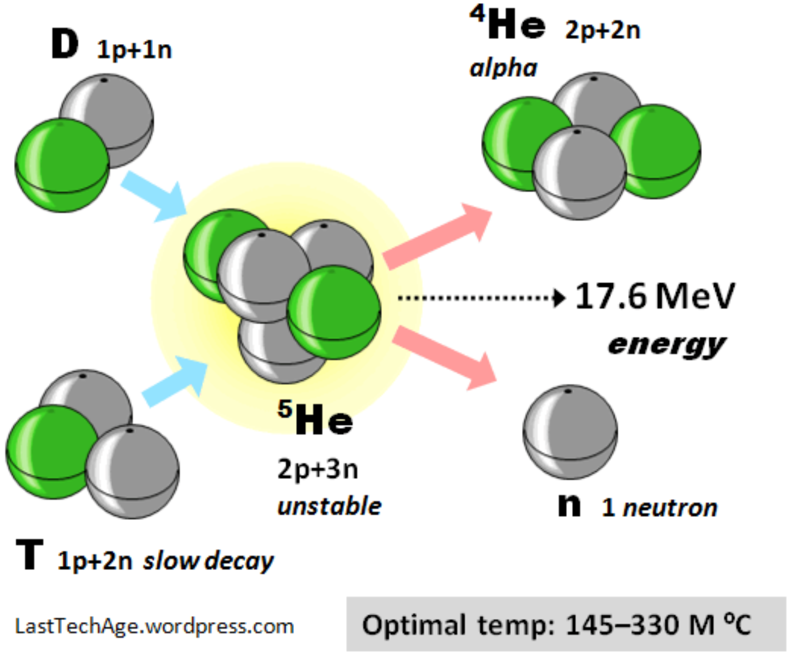


Idea: Carbon-free energy source for the 21st century and well beyond

Fusion energy in the laboratory



This process has by far the highest reaction rate under experimentally accessible conditions:



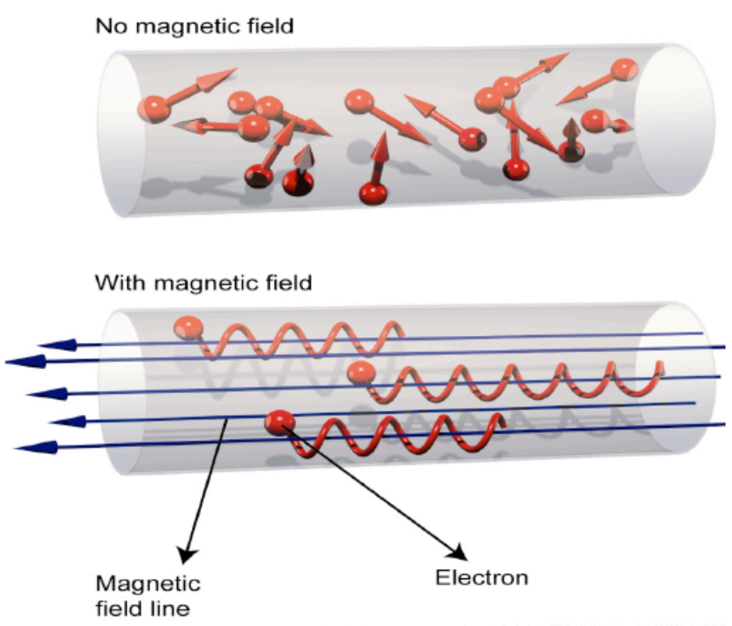
Still, temperatures of about 100 million degrees are required!
Thus, we are dealing with a fully ionized gas (plasma)

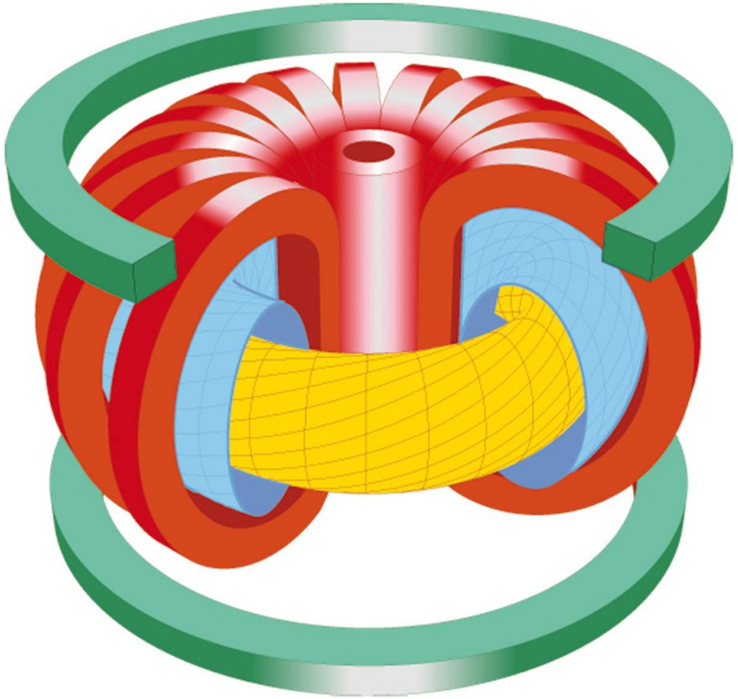
Magnetic confinement of fusion plasmas

Charged particles basically follow magnetic field lines

Helically twisted field lines span nested magnetic surfaces

Such an axisymmetric device is called “tokamak”



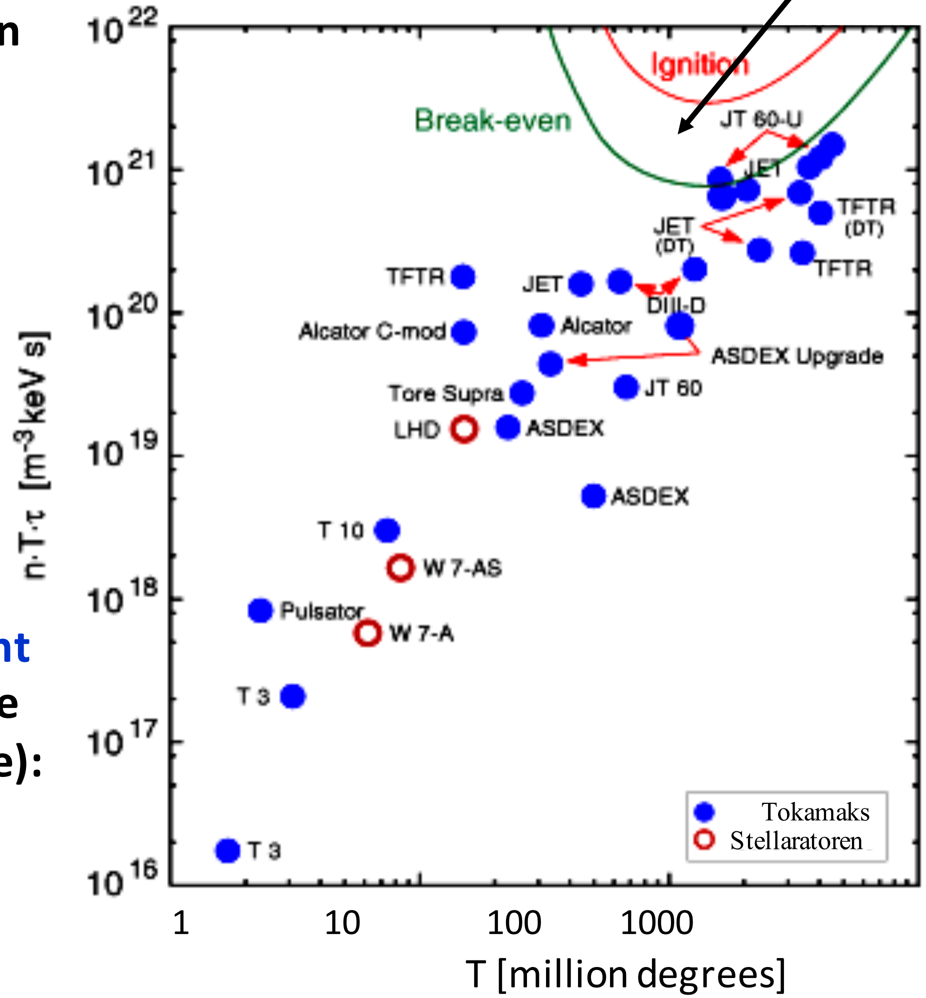


Fusion research: Towards burning plasmas

Self-heating must compensate energy losses:

- **Electromagnetic radiation**
- **Turbulent transport**

ITER

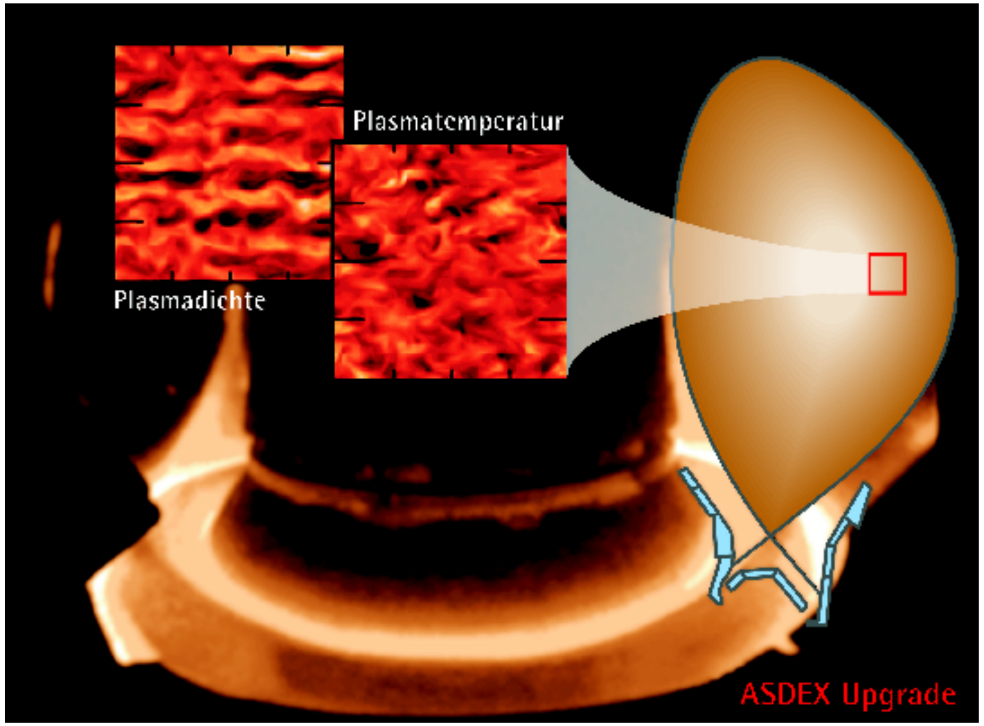


Key requirements:

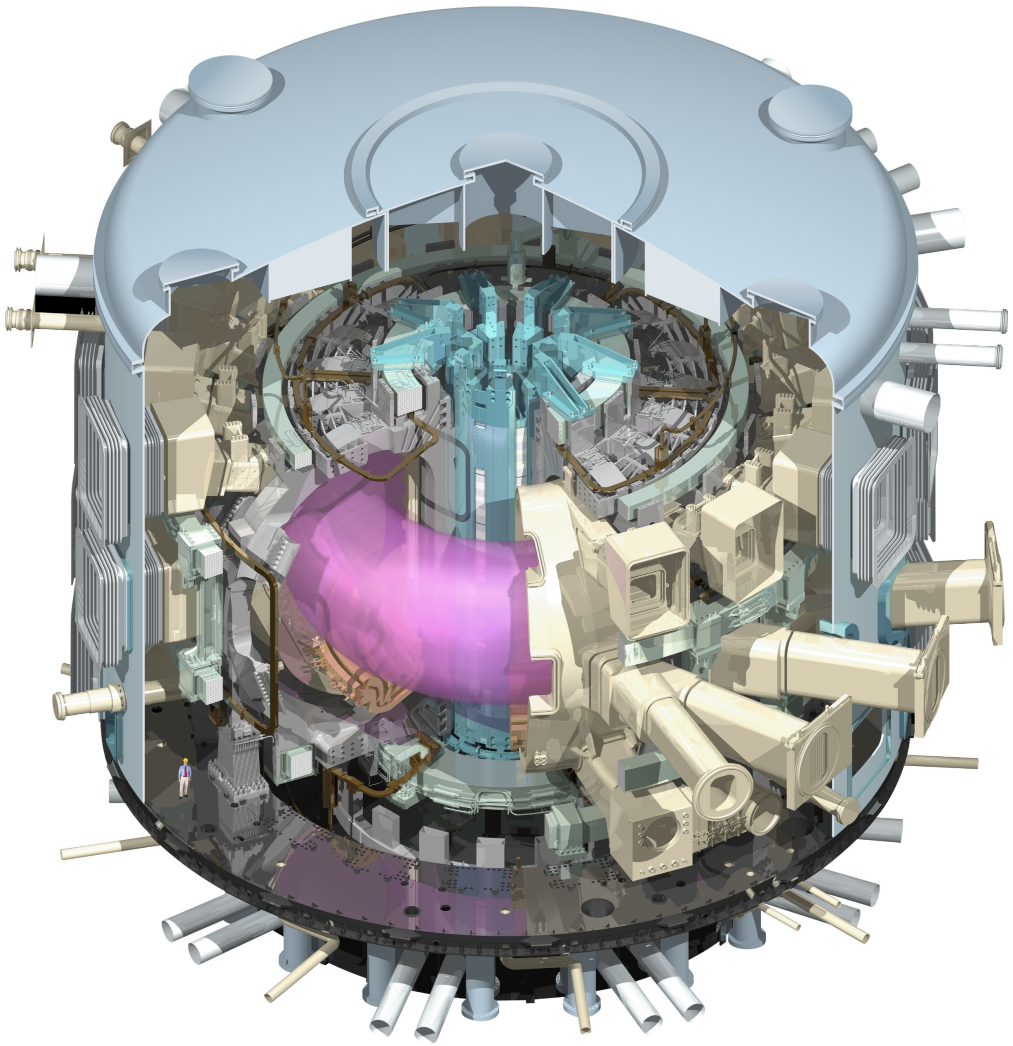
- **Large central pressure** **(limited by onset of large-scale instabilities)**
- **Large energy confinement time** **(limited by small-scale instabilities, i.e. turbulence):**

$$\tau_E = E_{plasma} / P_{loss}$$

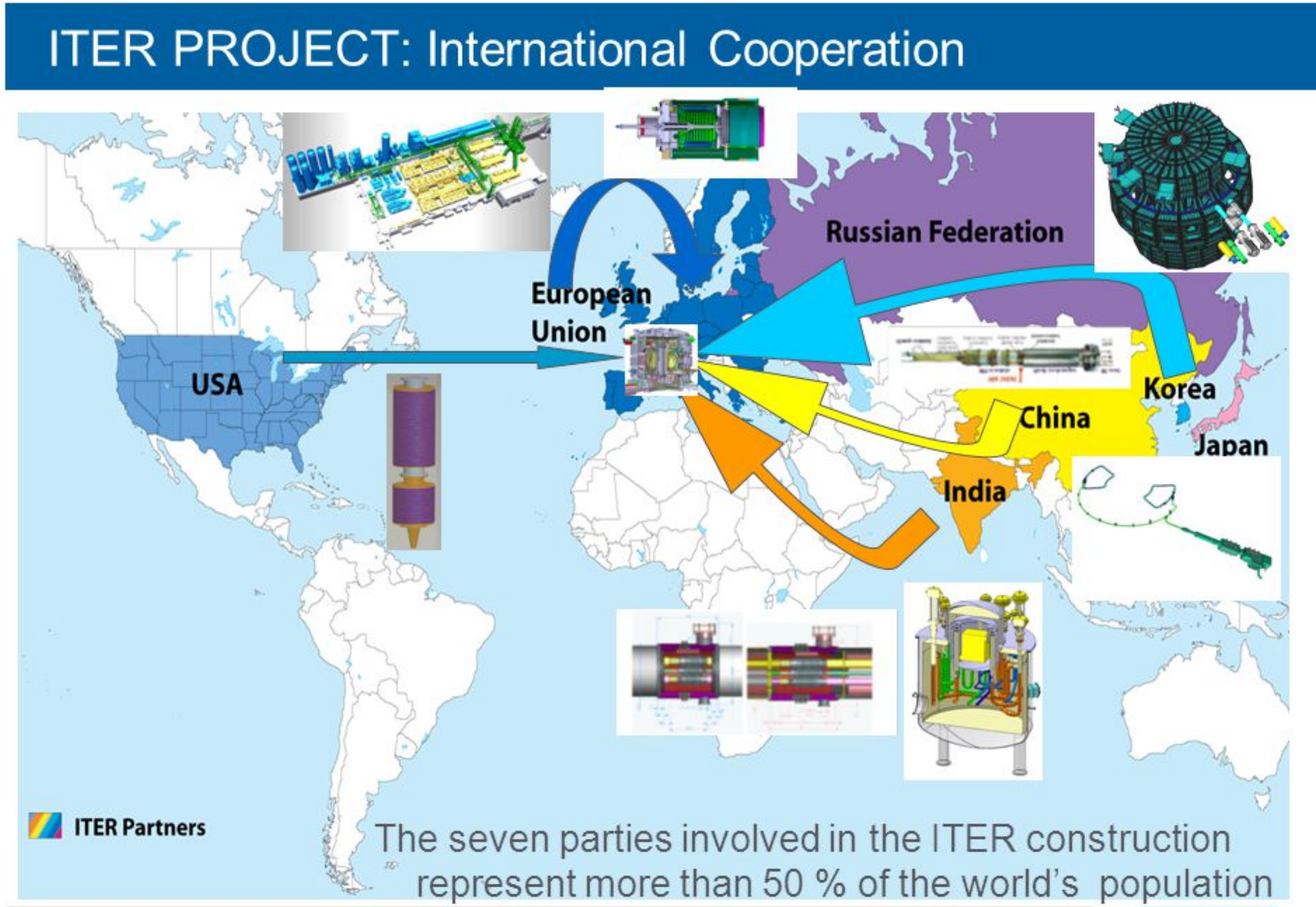
Radial heat transport due to small-scale turbulence controls energy confinement time



The international ITER project



Goal: 500 MW of fusion power www.iter.org



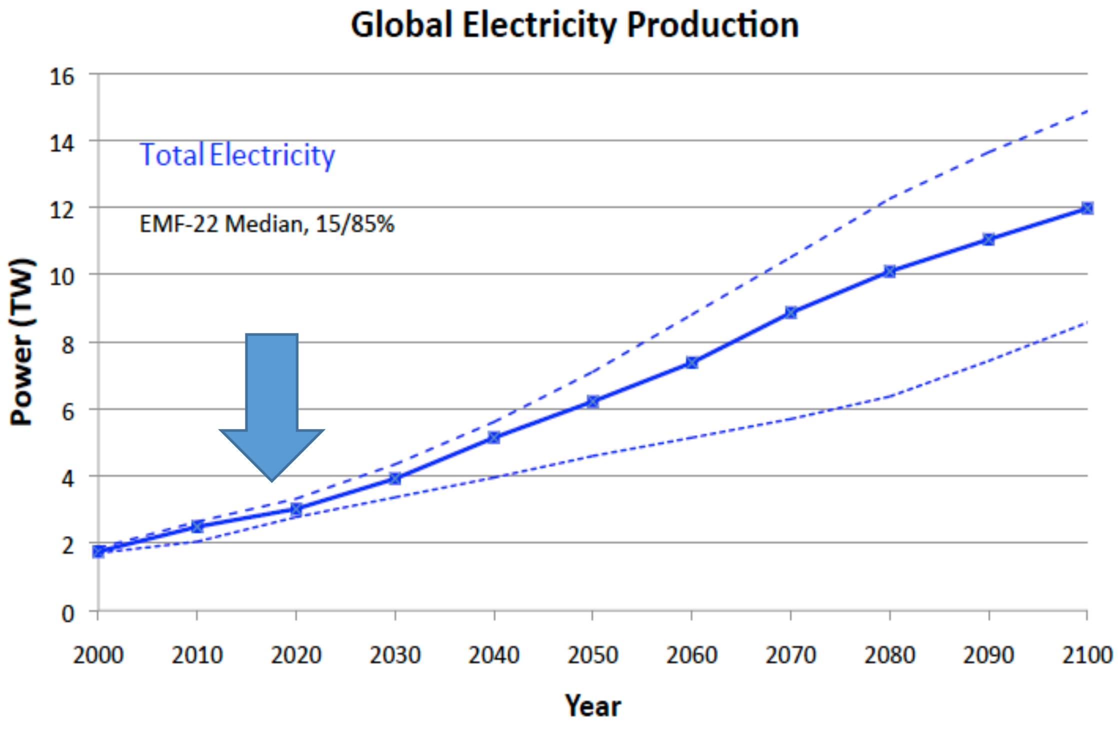
ITER construction site in Southern France

Resources are practically unlimited

Deuterium in a bath tub full of water and **Lithium** in a used laptop battery suffice for a family over 50 years

Global electricity needs will keep increasing

Energy Modeling Forum 22
100 models from 15 research groups (Clarke 2009)



From trial-and-error to predict-first: The role of High Performance Computing

At the forefront of supercomputing since the 70's

NERSC HISTORY

Powering Scientific Discovery Since 1974

Contact: Jon Bashor, jbashor@lbl.gov, +1 510 486 5849

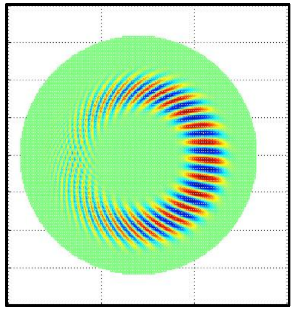
The oil crisis of 1973 did more than create long lines at the gas pumps — it jumpstarted a supercomputing revolution.

The quest for alternative energy sources led to increased funding for the Department of Energy's Magnetic Fusion Energy program, and simulating the behavior of plasma in a fusion reactor required a computer center dedicated to this purpose. Founded in 1974 at Lawrence Livermore National Laboratory, the Controlled Thermonuclear Research Computer Center was the first unclassified supercomputer center and was the model for those that followed.

Over the years the center's name was changed to the National Magnetic Fusion Energy Computer Center and later the National Energy Research Supercomputer Center (NERSC). In 1983 NERSC's role was expanded beyond the fusion program, and it began providing general computing services to all of the programs funded by the DOE Office of Energy Research (now the Office of Science). The current name was adopted in 1996 when NERSC relocated to Lawrence Berkeley National Laboratory and merged with Berkeley Lab's Computing Sciences program. The name change — from "Supercomputer Center" to "Scientific Computing Center" — signaled a new philosophy, one of making scientific computing more productive, not just providing supercomputer cycles.

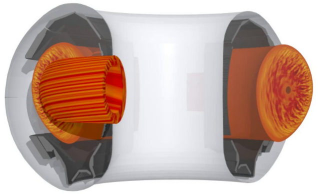
Towards a virtual fusion plasma

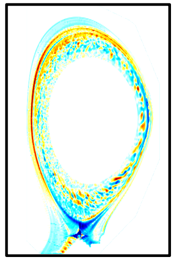
Increasing fidelity & modeling capability with increasing computing power



Gigaflops

Core: ion-scale electrostatic physics in simplified geometry

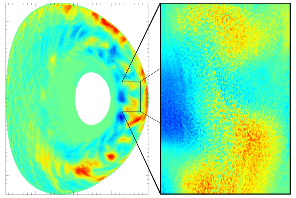


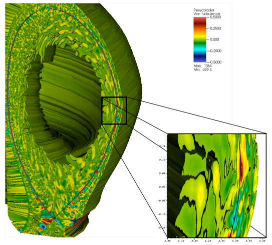


Teraflops

Core: adding kinetic electron electromagnetic physics in a torus

Edge: ion+neutral electrostatic physics in a torus

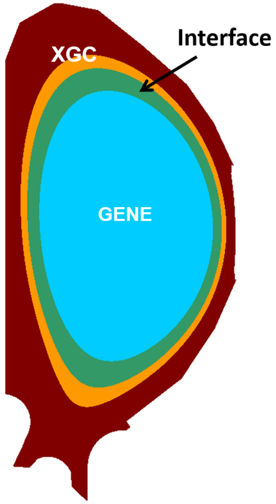




Petaflops

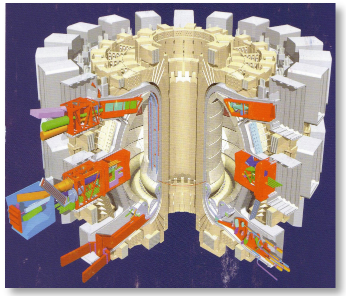
Core: adding electron-scale physics

Edge: adding kinetic electron electrostatic physics



Exaflops

Core-edge coupled studies of whole-device ITER, incl. turbulence, MHD instability, fast particles, heating, and plasma-wall interactions



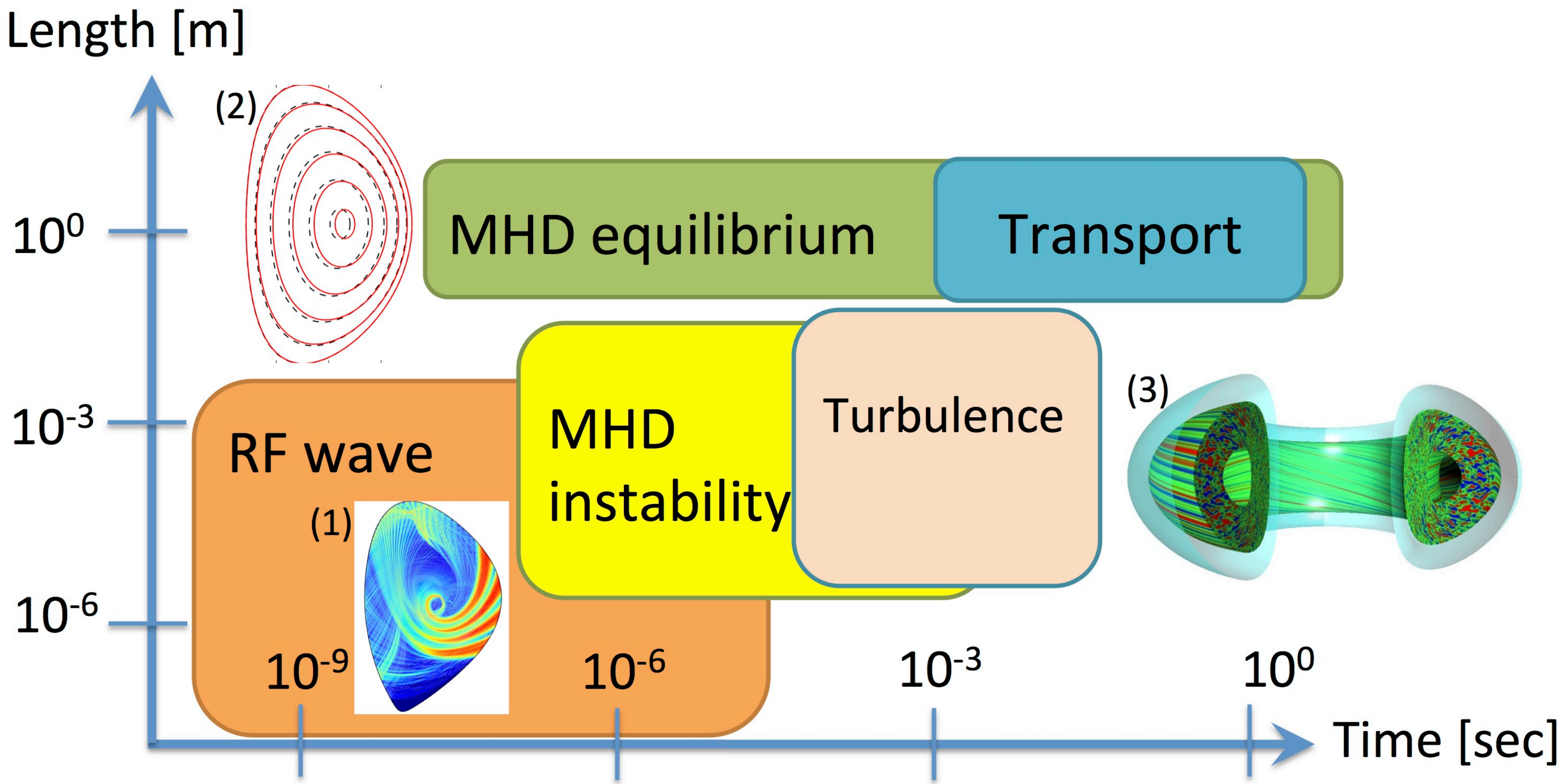
Beyond

Whole device modeling of all relevant fusion science

Acknowledgements: ECP

Goals: prepare and interpret **ITER** discharges, guide the development of **power plants**

The multiscale, multiphysics challenge

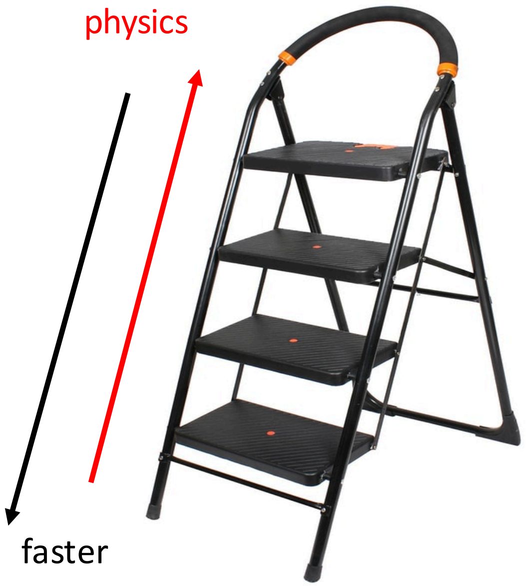


Many nonlinear interactions; we cannot use a simple "superposition principle"

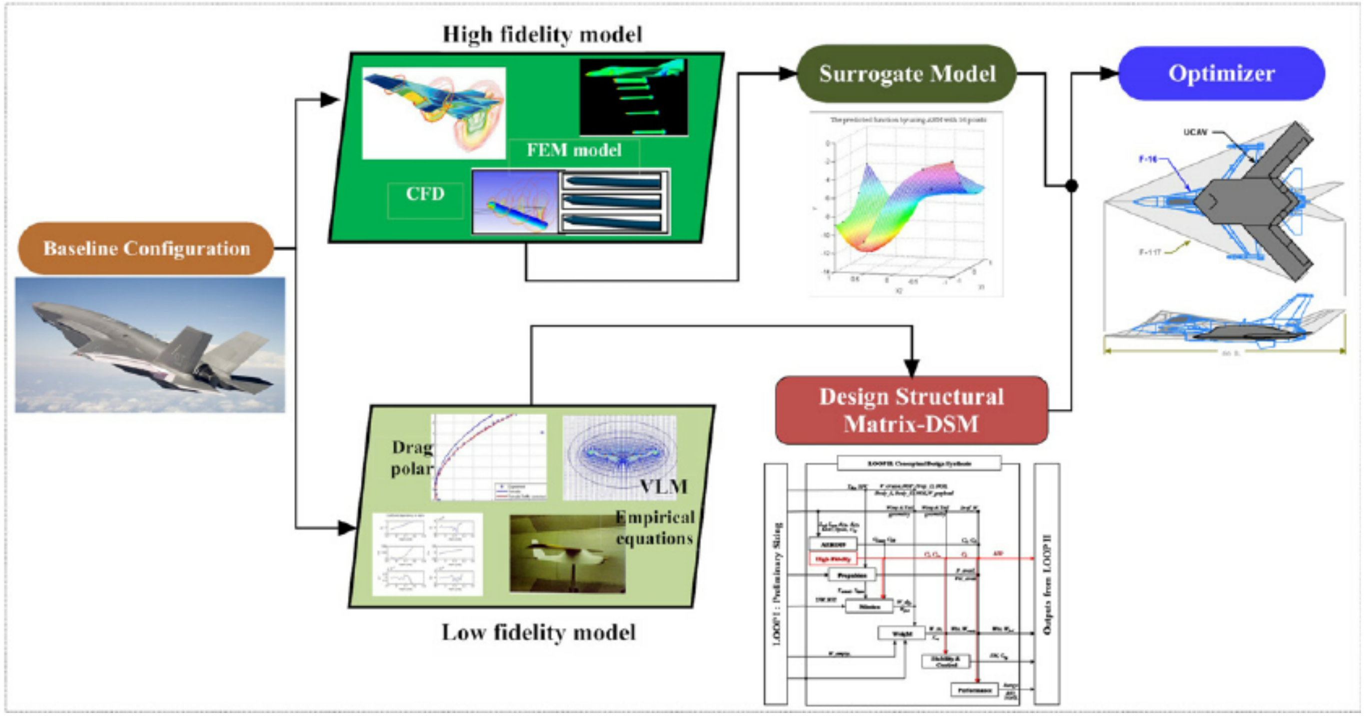
A multi-fidelity approach

An example:

more complete physics



faster



Talks by K. Willcox & B. Peherstorfer

- High-fidelity models provide reliable **predictive capability**
- Lower-fidelity models foster **high-throughput computing**
- Both are needed – together

A high-fidelity model for determining turbulent transport (i.e., the energy confinement time): The GENE code

Global Gyrokinetic Simulation of Turbulence in

ASDEX Upgrade

GENE

gene.rzg.mpg.de

Fluid models don't work – use (gyro-)kinetics!

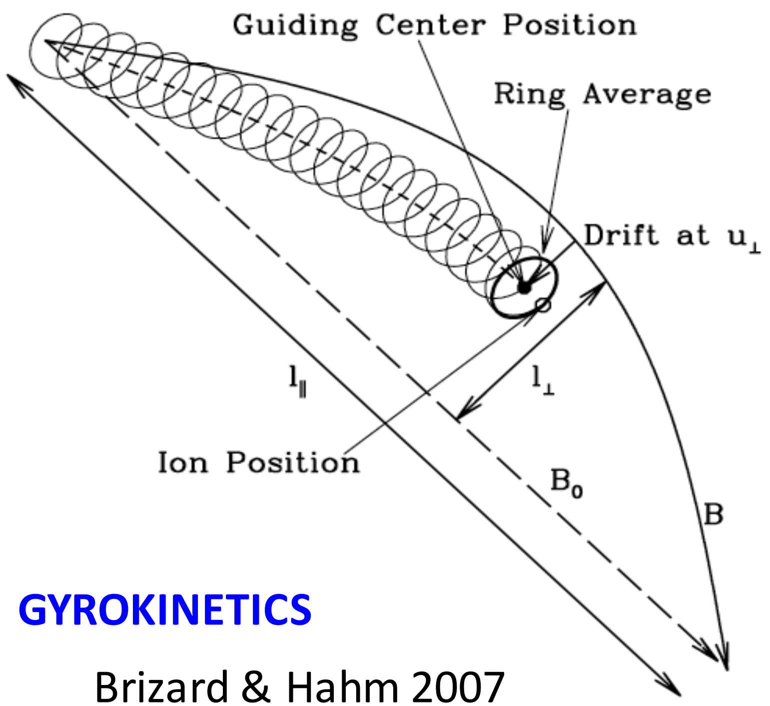
Hot and/or dilute plasmas are only weakly collisional: 6D Vlasov-Maxwell equations

$$\frac{\partial f_\alpha}{\partial t} + \mathbf{v} \cdot \nabla f_\alpha + \frac{q_\alpha}{m_\alpha}\left[\mathbf{E} + \frac{\mathbf{v} \times \mathbf{B}}{c}\right] \cdot \nabla_v f_\alpha = 0$$

α = particle species

$$f_\alpha = f_\alpha(\mathbf{x}, \mathbf{v}, t)$$

...from the Liouville equation via the BBGKY hierarchy



GYROKINETICS

Brizard & Hahm 2007

Strong background magnetic field:

Eliminate fast gyromotion; consider slow dynamics of **guiding centers**

$$f = f(\mathbf{X}, v_\parallel, \mu; t)$$

$$\boxed{\frac{\partial f}{\partial t} + \dot{\mathbf{X}} \cdot \frac{\partial f}{\partial \mathbf{X}} + \dot{v}_\parallel \frac{\partial f}{\partial v_\parallel} = 0}$$

Reduction of effort by ~12 orders of magnitude (elimination of irrelevant spatio-temporal scales & reduction from 6D to 5D)

Additional gain from using field-aligned coordinates

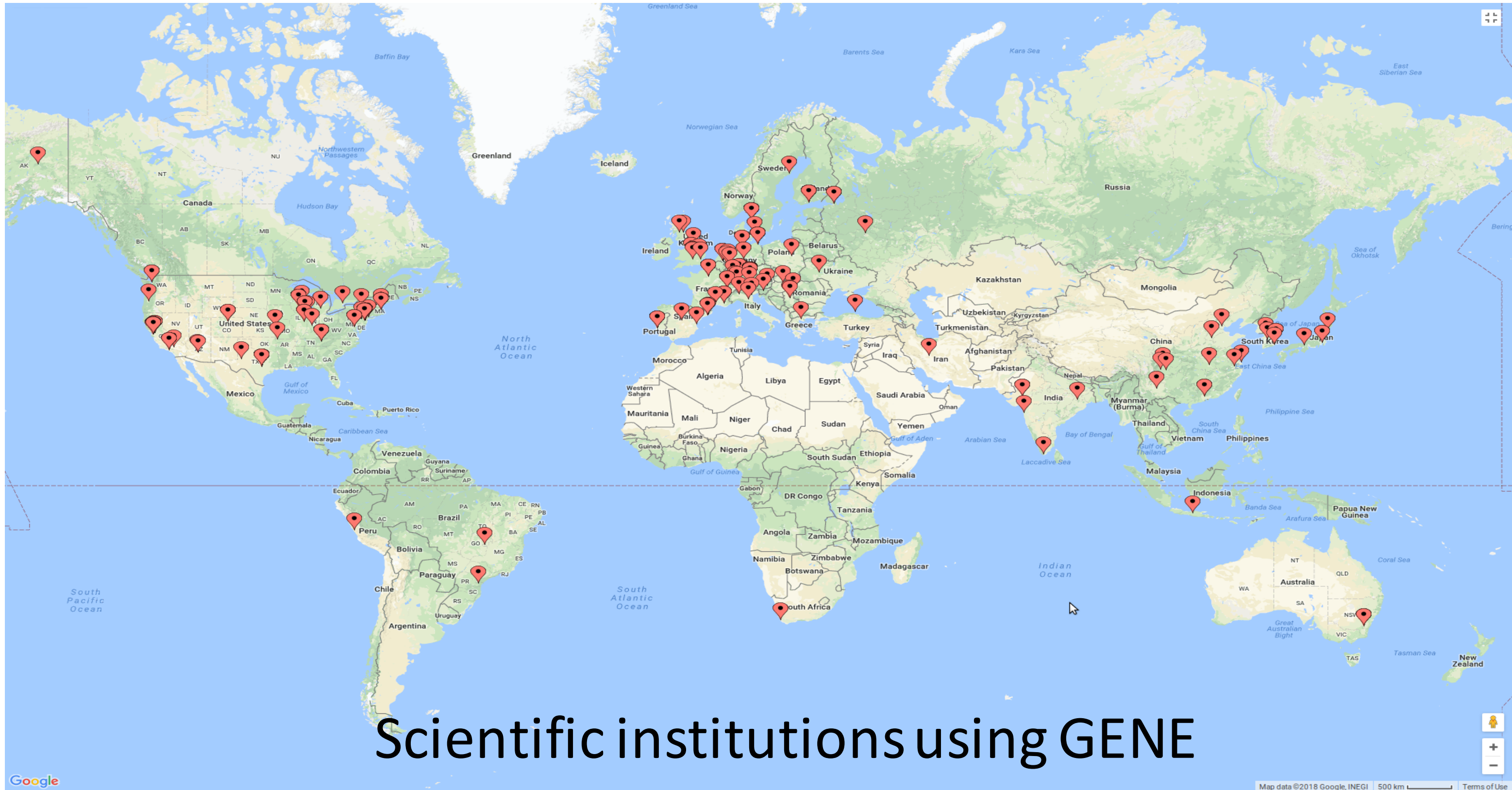
The gyrokinetic code GENE

- First GENE publication: Jenko et al., Physics of Plasmas 2000 (~600 citations)
- More than 100,000 lines of source code, plus 200,000 lines for pre-/post-processing
- Part of **PRACE's Unified European Applications Benchmark Suite**

- Open source policy
- World-wide user base:
  genecode.org

Part of an ecosystem of codes for fusion research

Also used in astrophysics

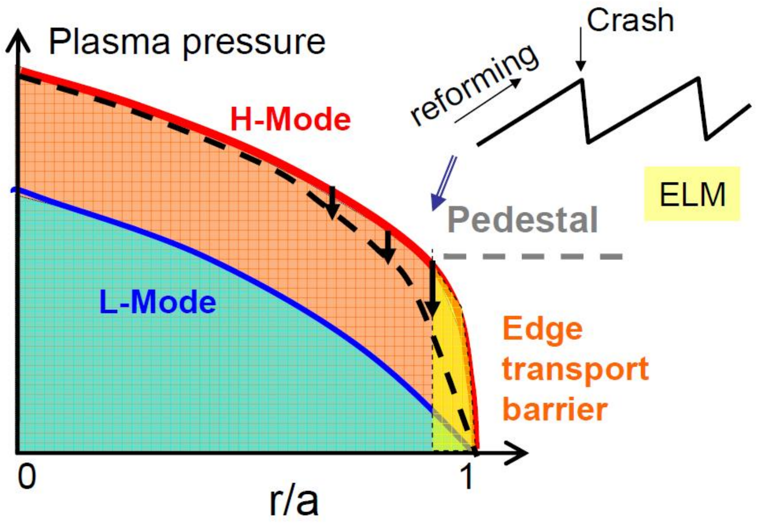


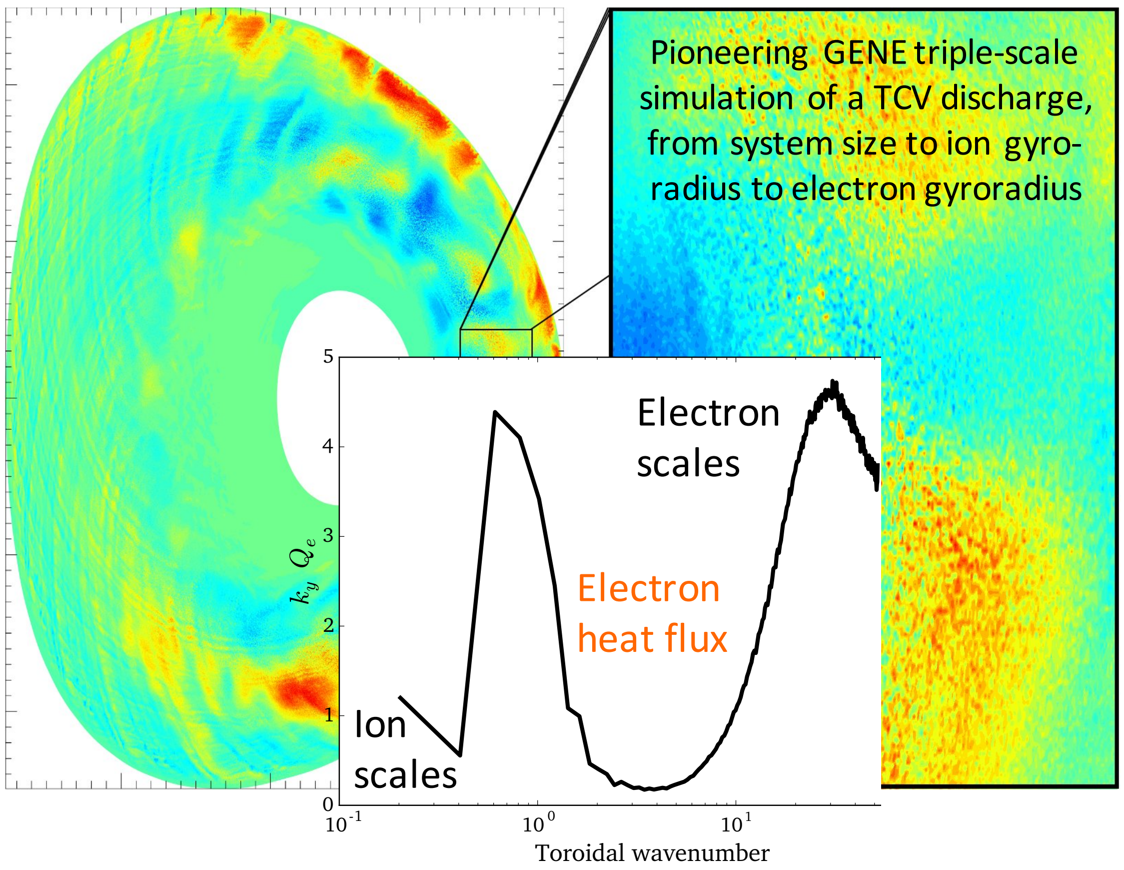
Scientific institutions using GENE

Science highlights (just two examples)

Prediction of relevant contributions to turbulent transport at very small – hiterto neglected – spatio-temporal scales (experimentally confirmed)

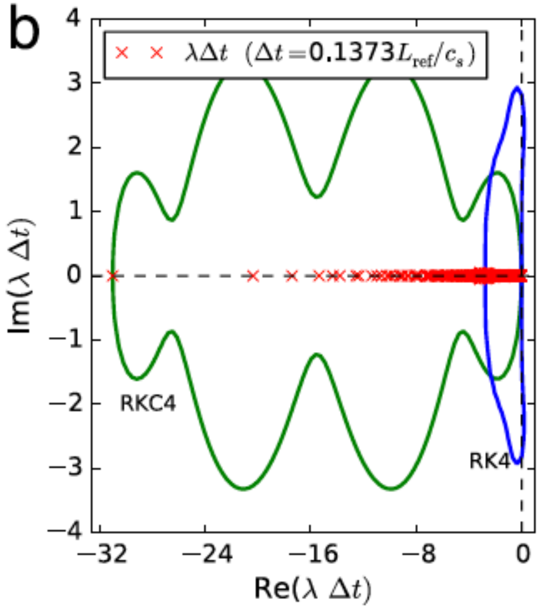
Indications that the physics of the "pedestal region" in ITER may differ from that in present-day devices





Some background on GENE

- The underlying nonlinear PDEs are discretized on a fixed grid in 5D phase space
- Apply CFD-type (mix of spectral, finite difference, and finite volume) methods
- Explicit time-stepping facilitates scalability
- Time step is maximized during initialization



Runge–Kutta–Chebychev schemes

Doerk & Jenko CPC 2014

- 5D domain decomposition
- Primarily pure MPI or hybrid MPI/OpenACC programming model
- Auto-tuning (optimal subroutines and processor layout)

Some software engineering aspects

Version control system git https://github.com/git/git

Repository hosted by Max Planck Computing & Data Facility

Web interface installed at MPCDF – contains wiki, CI/CD, interface to bug tracker

Development repository managed via GitLab https://gitlab.com

Regression tests are performed automatically once per day/week

User repository at https://gitta.rzg.mpg.de/GENE

release tags and master branch mirrored from development repository

Source code documentation via Doxygen and FORD (FORtran Documenter)

Porting GENE to Titan (@ORNL)

MPCDF Max Planck Computing & Data Facility
Tilman Dannert

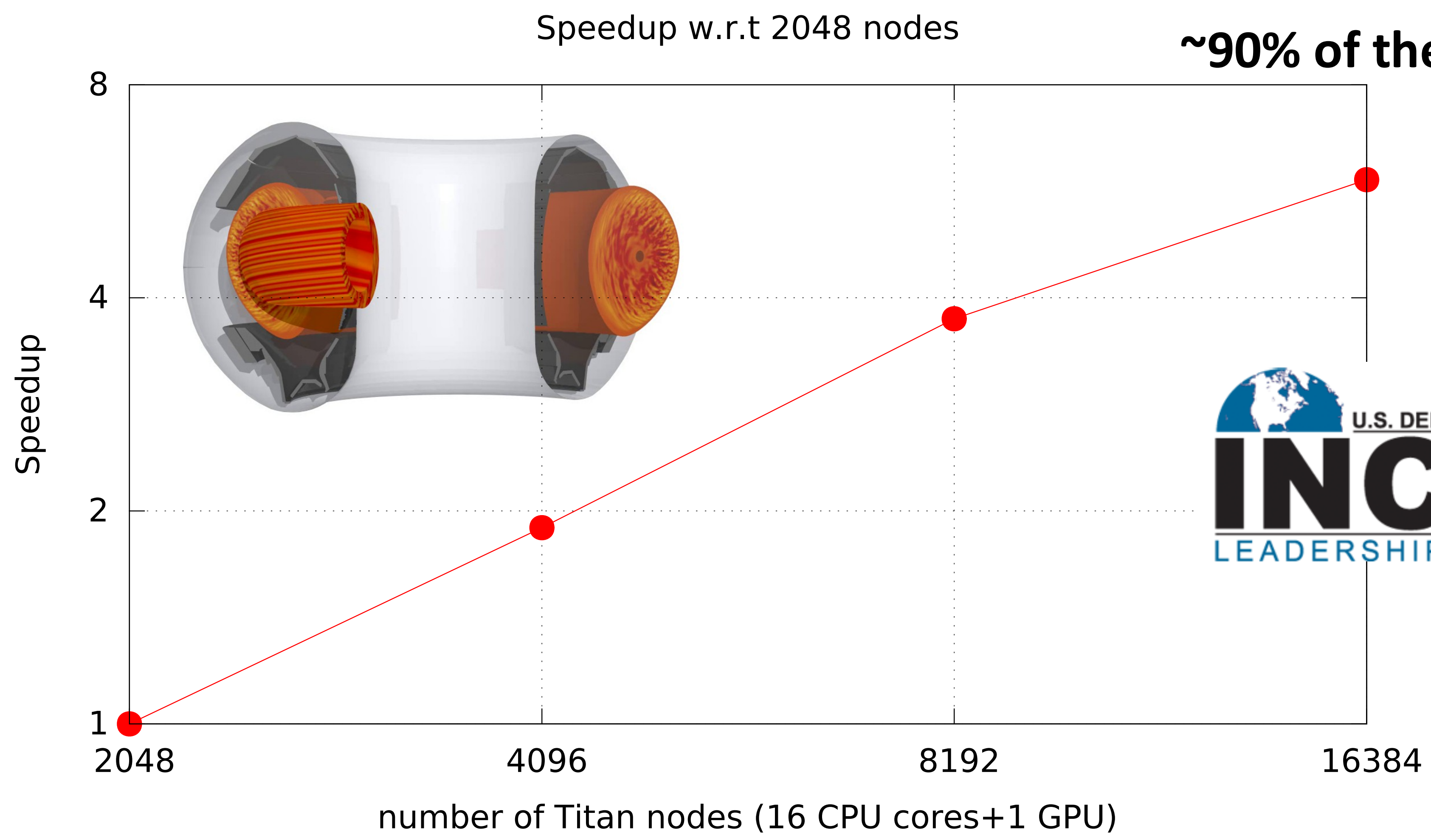
- 18,688 compute nodes
- Node: 16 Opteron cores + Kepler GPU
- Cray system

- First, the nonlinear terms were ported to GPUs using CUDA and cuFFT
- The bottleneck at this time was the data transfer to and from the GPUs
- Idea: Shift more computations to the GPUs; compute *all* terms there
- For maintenance reasons, we decided to use OpenACC instead of CUDA
- This lead us to use the PGI compiler

Strong scaling of GENE on Titan

Tilman Dannert

~90% of the machine



U.S. DEPARTMENT OF ENERGY INCITE LEADERSHIP COMPUTING

Optimizing large-scale codes (like GENE) on pre-exascale systems may take <u>person-years</u>

From Titan to Summit (2018)

2,282,544 cores
187 PF (peak)

IBM's Power9 processors
NVIDIA's Volta GV100 GPUs

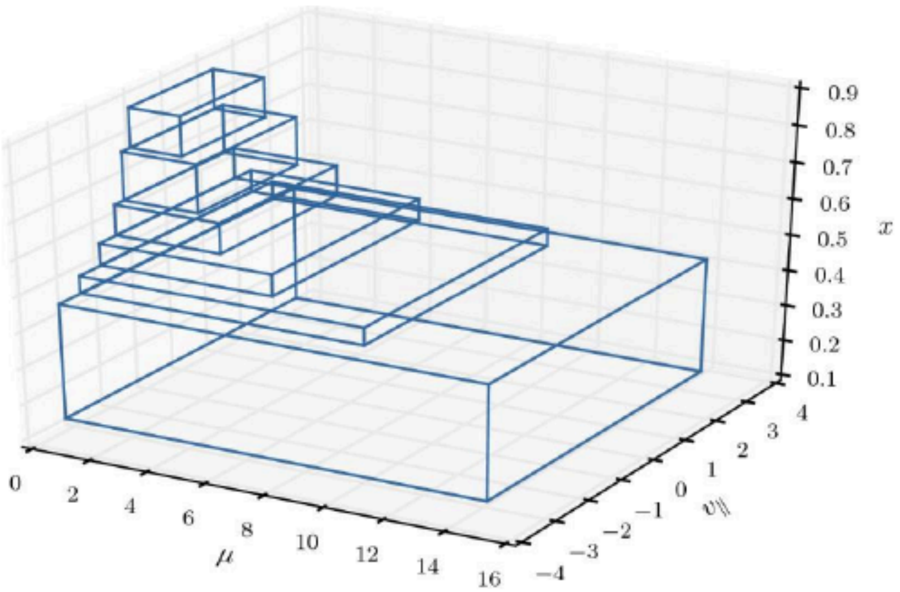
Much faster data motion

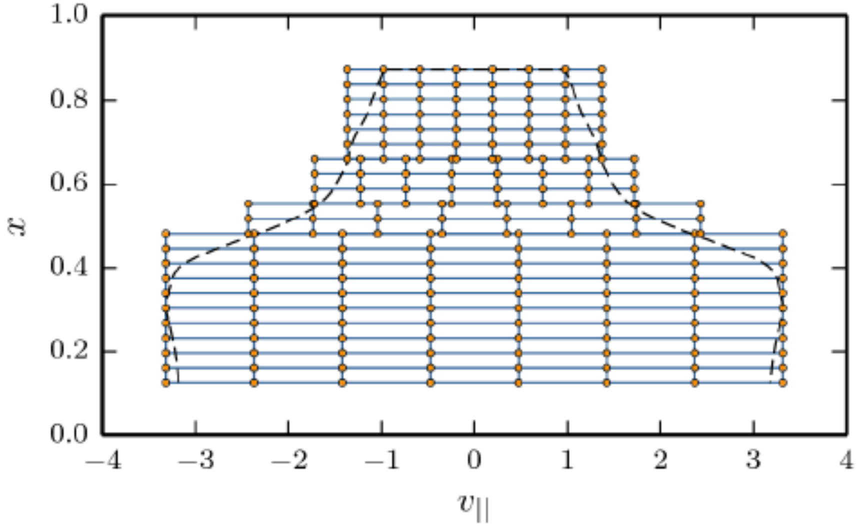
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband (/system/179397), IBM DOE/SC/Oak Ridge National Laboratory (/site/48553) United States	2,282,544	122,300.0	187,659.3	8,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway	10,649,600	93,014.6	125,435.9	15,371

Most powerful supercomputer in the world
(Top 500 List, June 2018)

Beyond brute force:
Role of applied math & computer science

Block-structured grids in velocity space



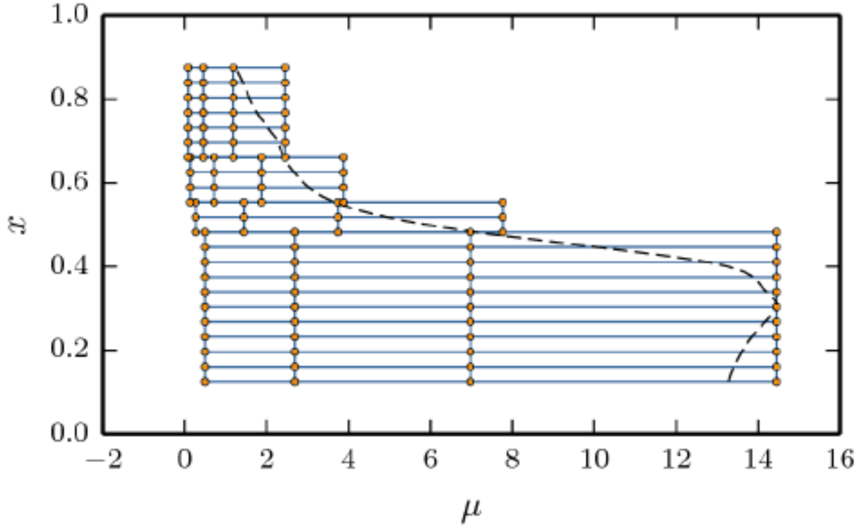


Effort:
~4 person-years

Gain:
~10x

Benefits:

- Preserves governing equation
- Reuses implementation for original regular grids
- Allows simultaneous development of physical model and computational grid
- For already five-six blocks, reduction in grid points equivalent to scaling approach



Jarema+ CPC 2016
Jarema+ CPC 2017

Sparse grid techniques

Cartesian grid

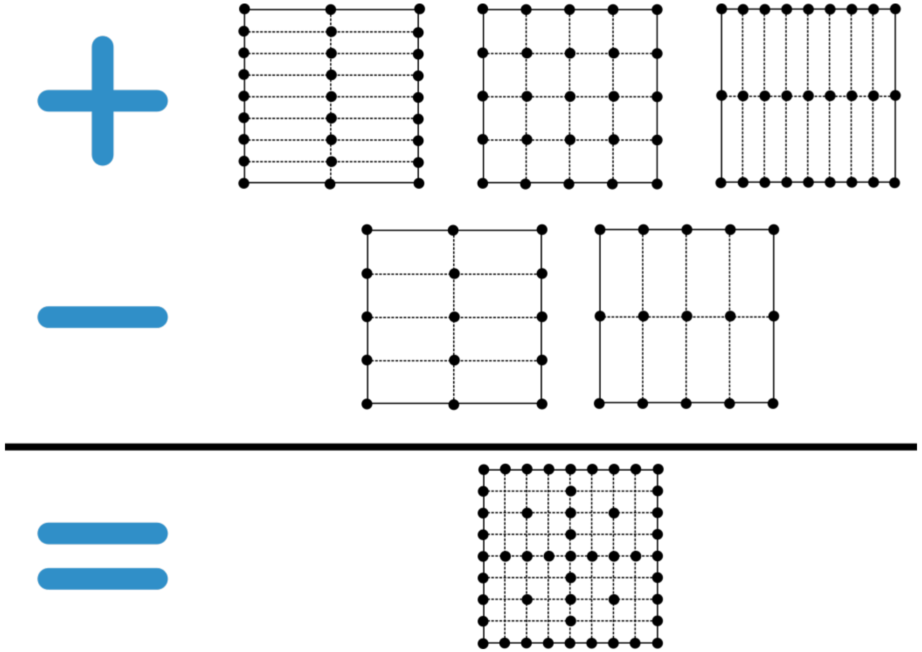
- Regular data structure
- Huge number of grid points in high dimensions
  "curse of dimensionality"

Resolution: 33 grid points per dimension	2D	5D
Cartesian grid	1,089	39,135,393
Combination tech.	641	206,358

Sparse grid combination technique

- Good approximation of the Cartesian grid solution
- Existing code (GENE) can be used more or less as it is
- **A new level of parallelism**
- **Algorithmic fault tolerance**

Talks by H. Bungartz and D. Pflüger



An ambitious project: Coupling two high-fidelity codes to create a whole device model

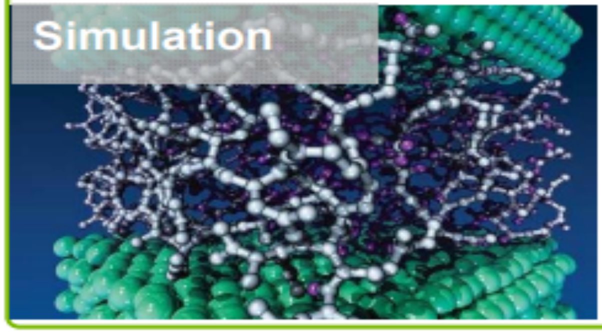
1st U.S. exascale system (2021)

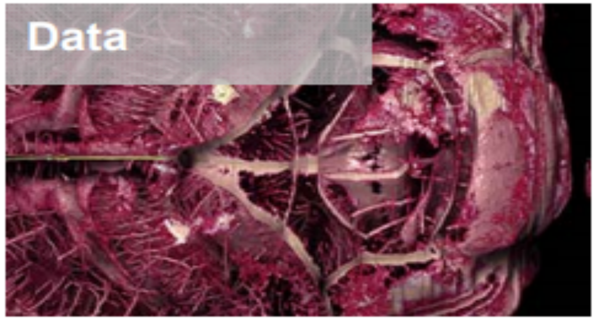
ALCF 2021 EXASCALE SUPERCOMPUTER – A21

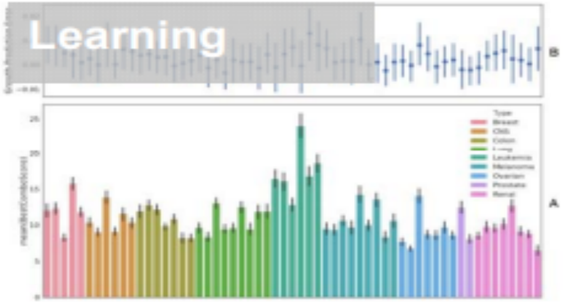
Intel/Cray Aurora supercomputer planned for 2018 shifted to 2021

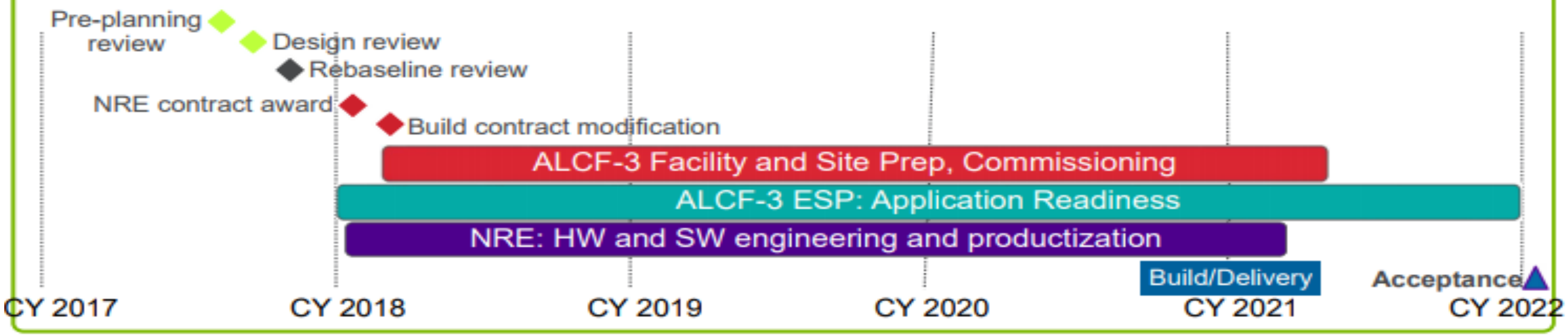
Scaled up from **180 PF to over 1000 PF**

Support for three “pillars”





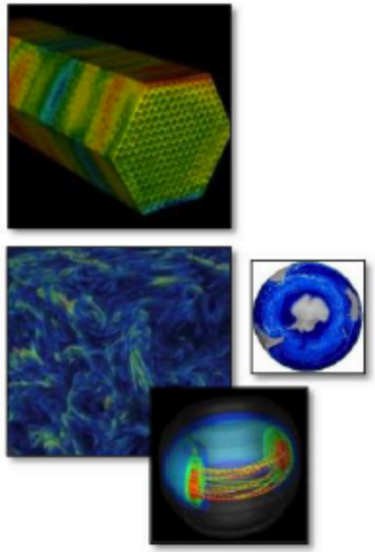


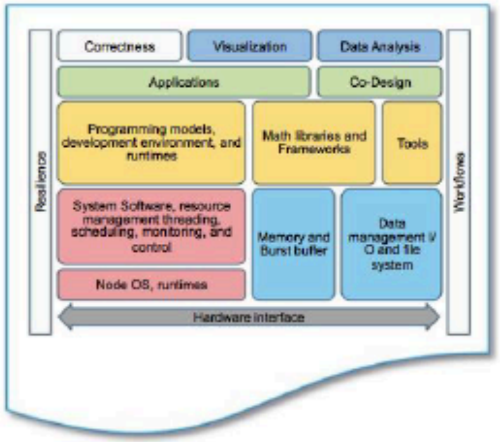


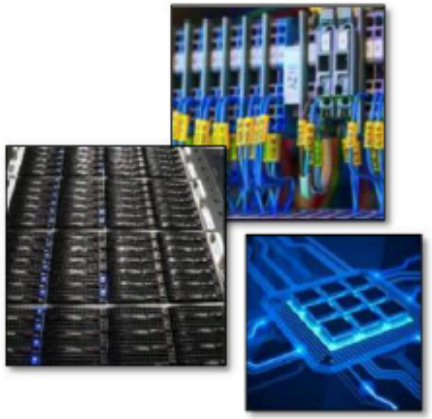
U.S. DOE Exascale Computing Project (ECP)

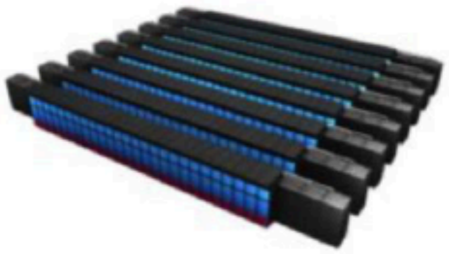
A holistic approach

Application Development	Software Technology	Hardware Technology	Exascale Systems
Science and mission applications	Scalable software stack	Hardware technology elements	Integrated exascale supercomputers







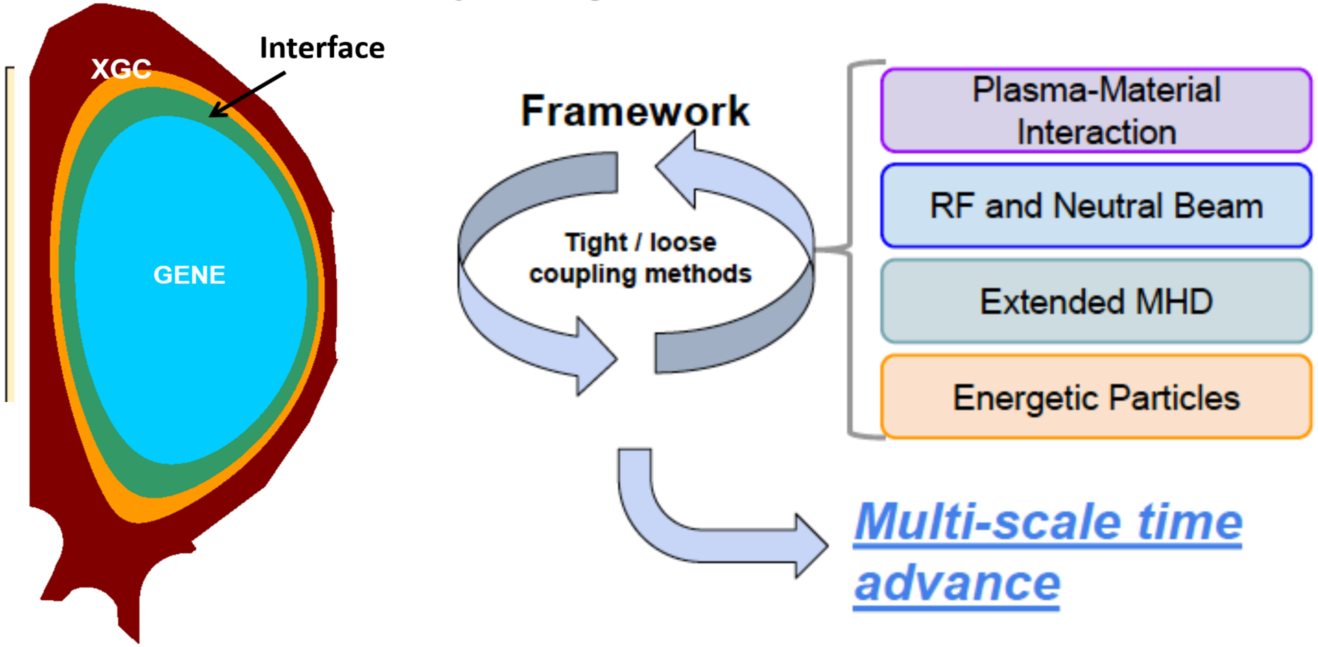


The ECP is a 7-year project with a cost range of $3.5B – $5.7B

Fusion Energy Application Development (2016-)

10-year Goal: A First-Principles-Based Whole Device Model that Covers the Full Space/Time Scales of a Reactor

- XGC full-f particle-in-cell technique with continuity across separatrix
- GENE continuum delta-f capability for core



Boulder

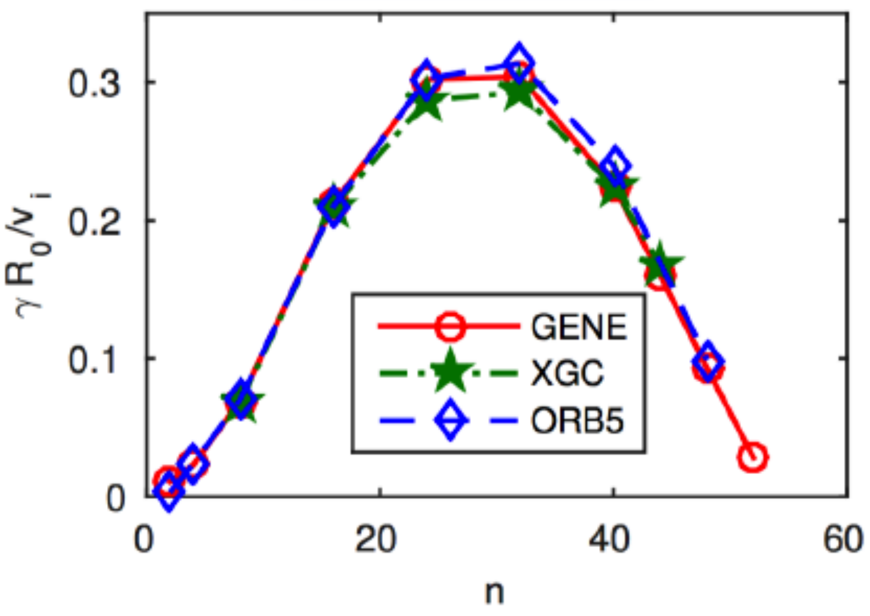
PPPL UCLA RUTGERS

EXASCALE COMPUTING PROJECT

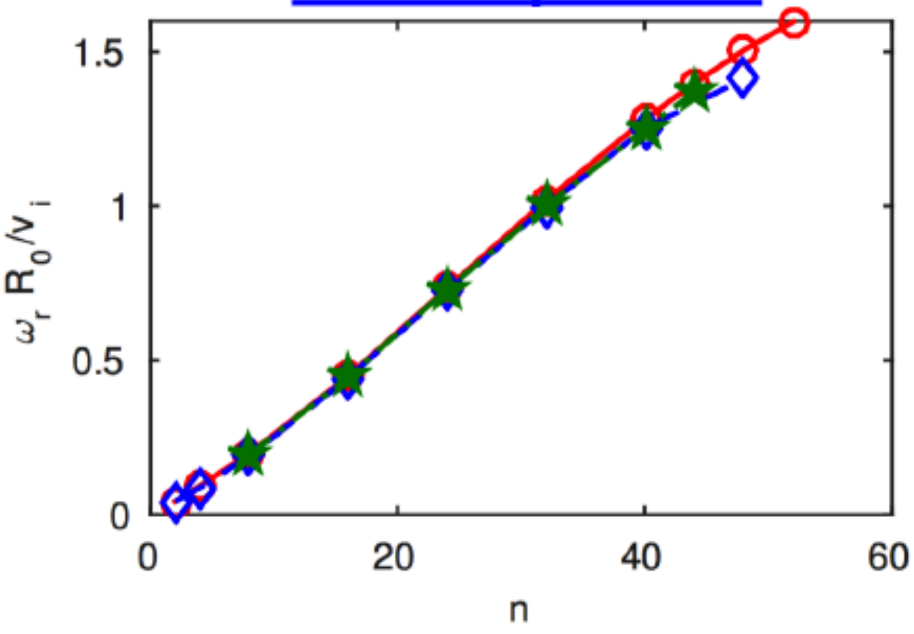
1st step: Code benchmarking

Cross-verification between GENE, XGC, and ORB5: bonus!
linear ITG instability (S. Ku, G. Merlo, E. Lanti (SPC, EPFL))

Growth rates



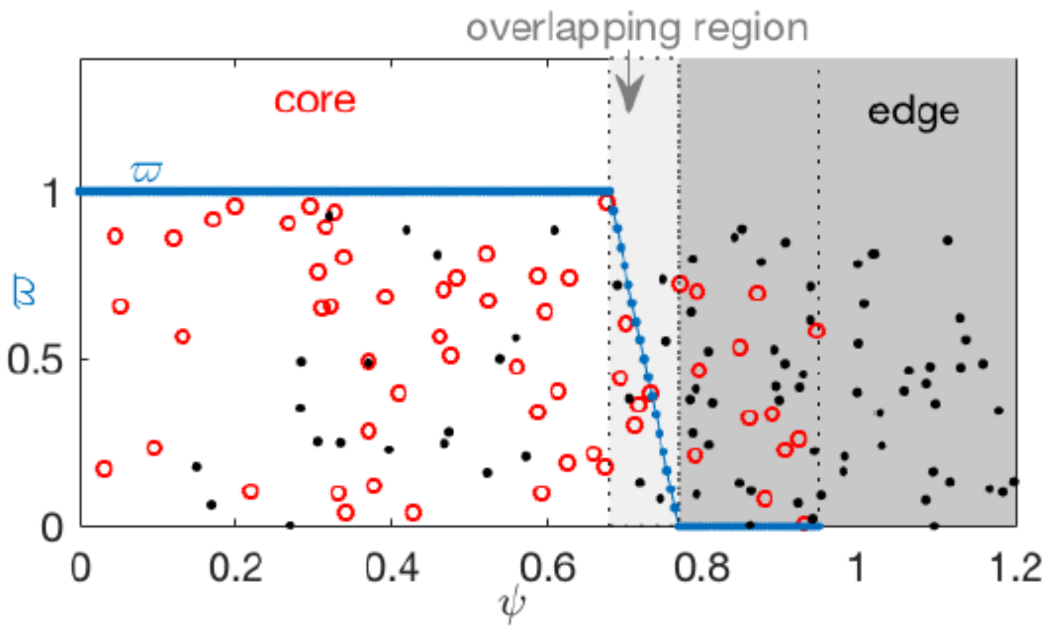
Real frequencies

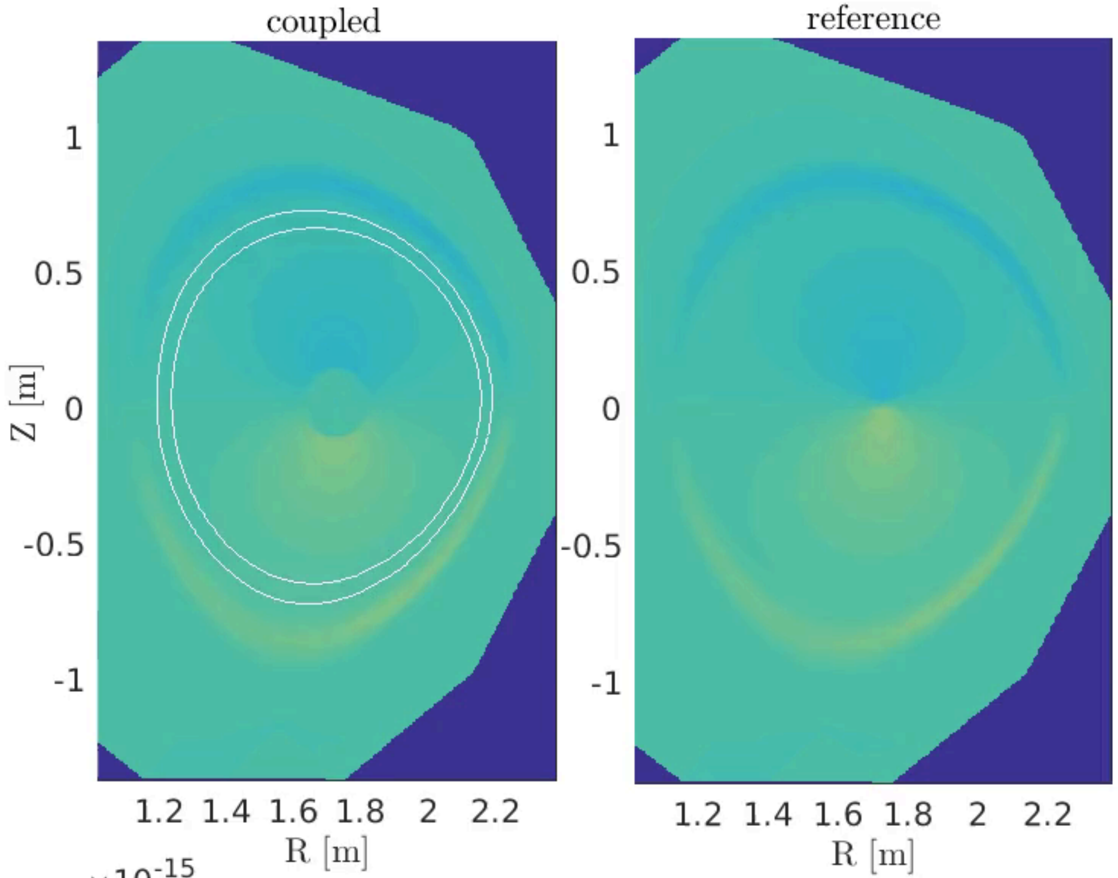


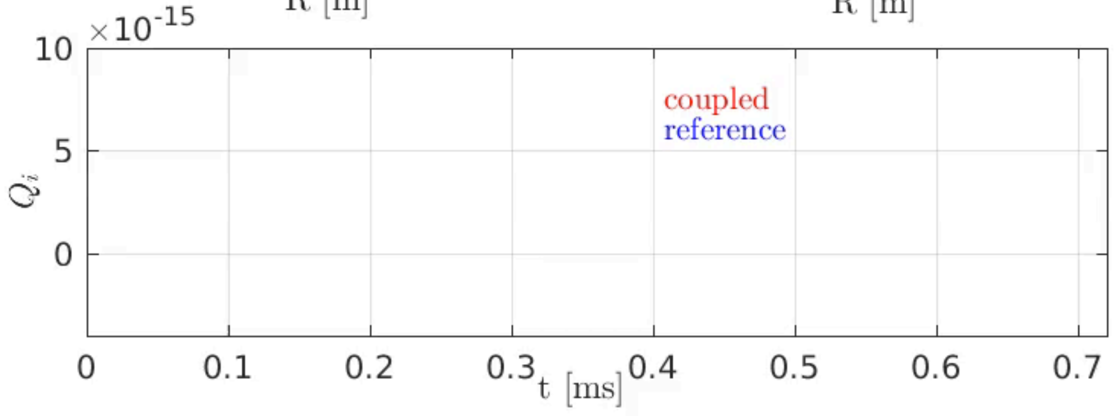
- Codes agree within 10% for all modes considered

2^{nd} step: XGC self-coupling

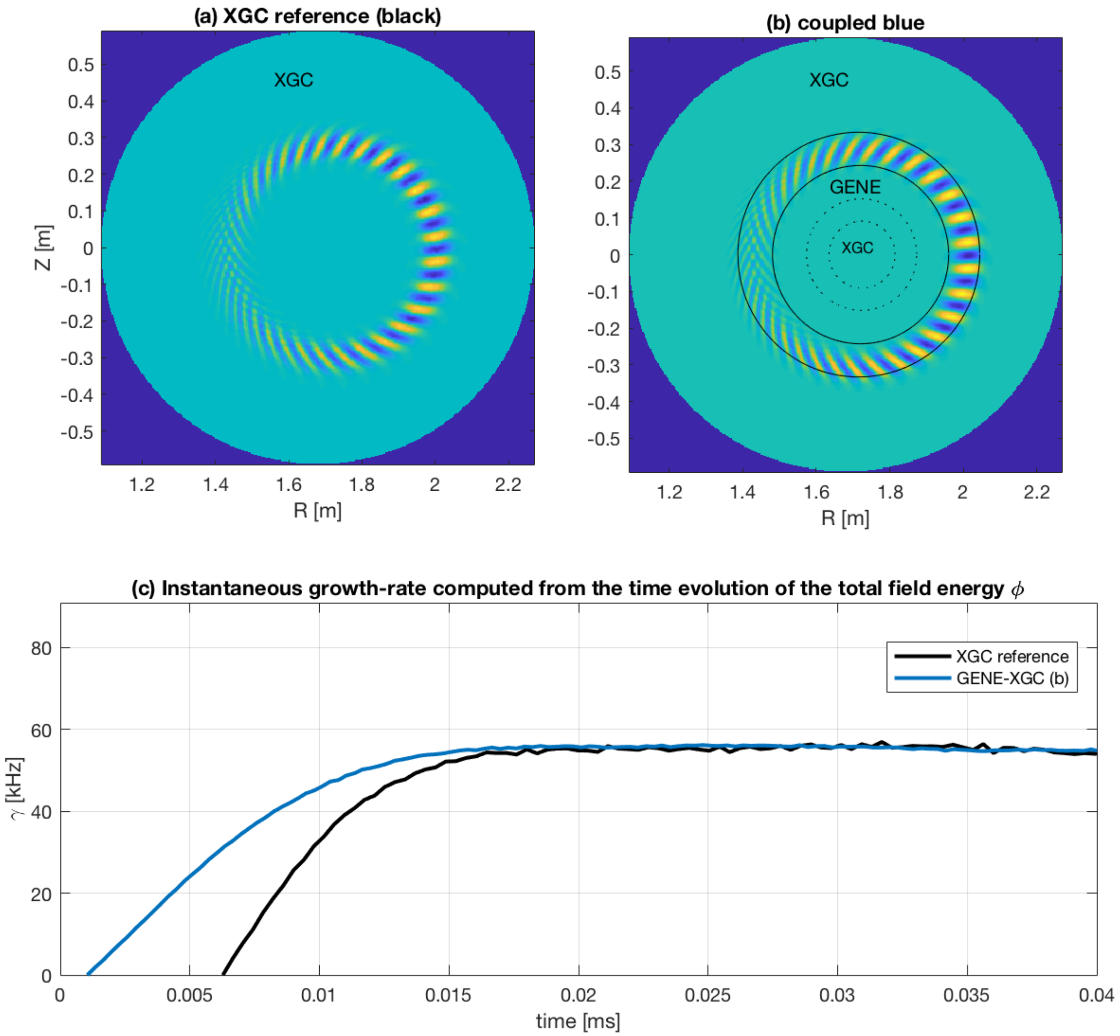
Global field solve via combined f

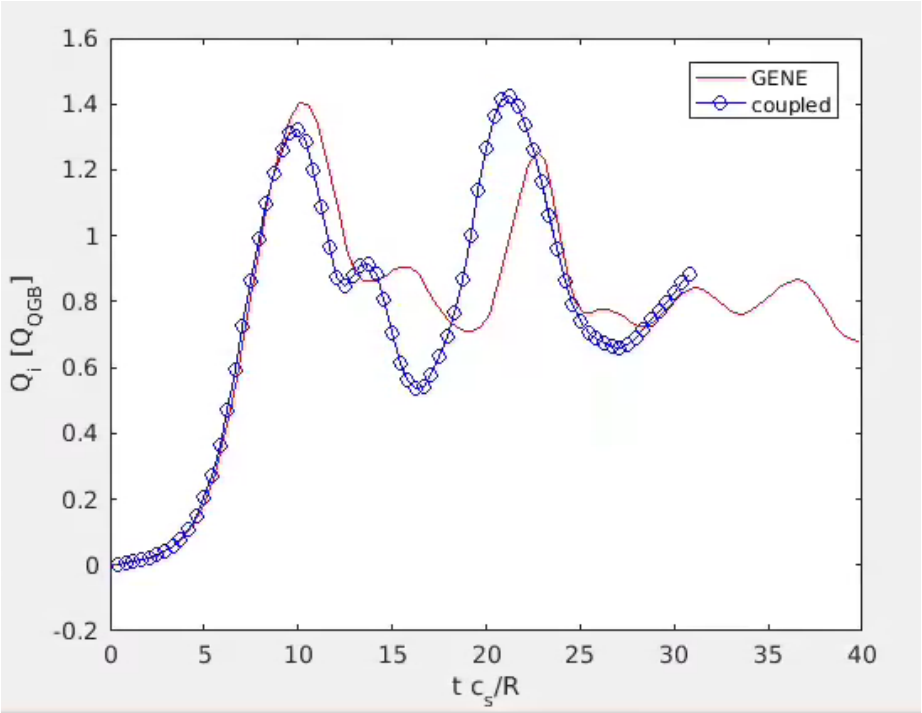






3rd step: GENE-XGC coupling



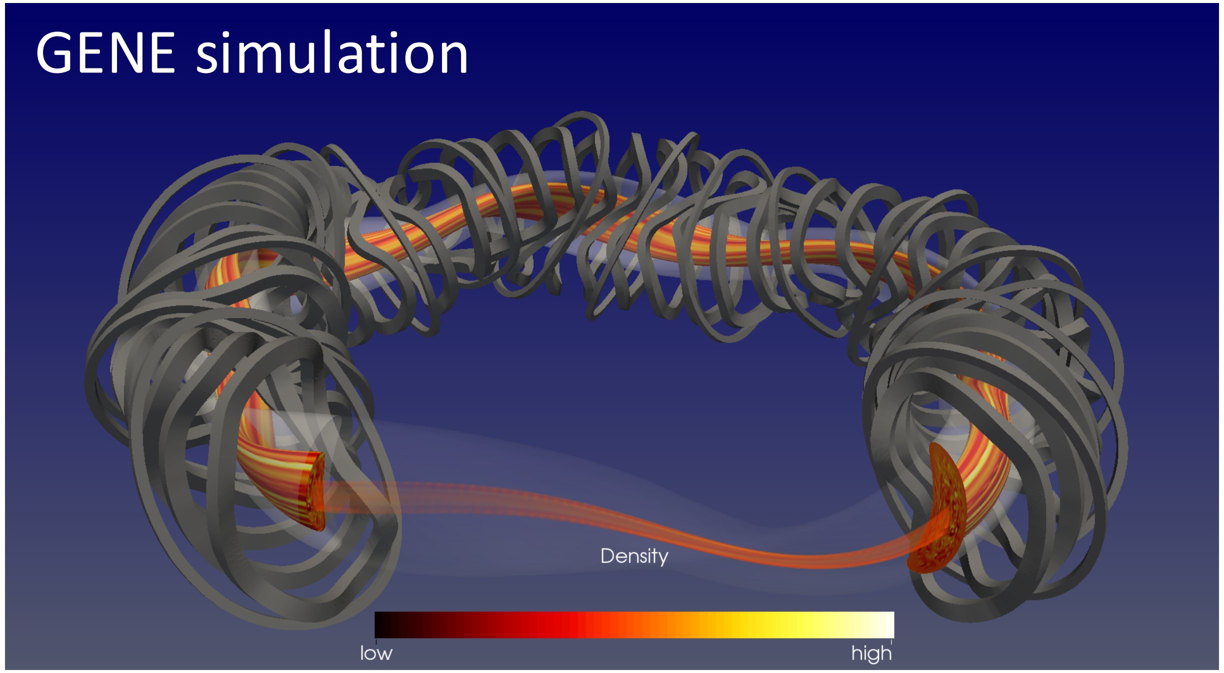


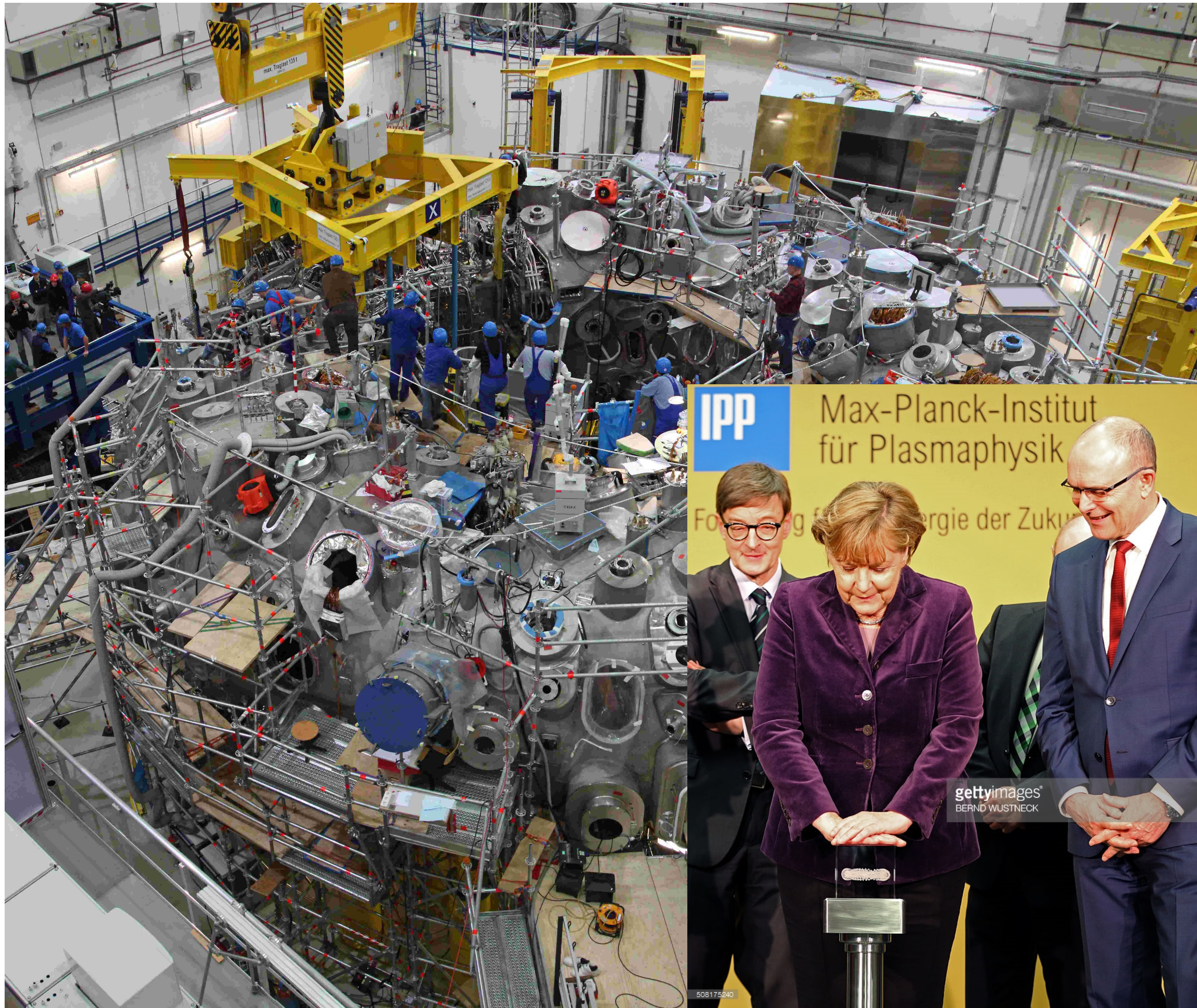
Optimizing the design of future fusion experiments on supercomputers

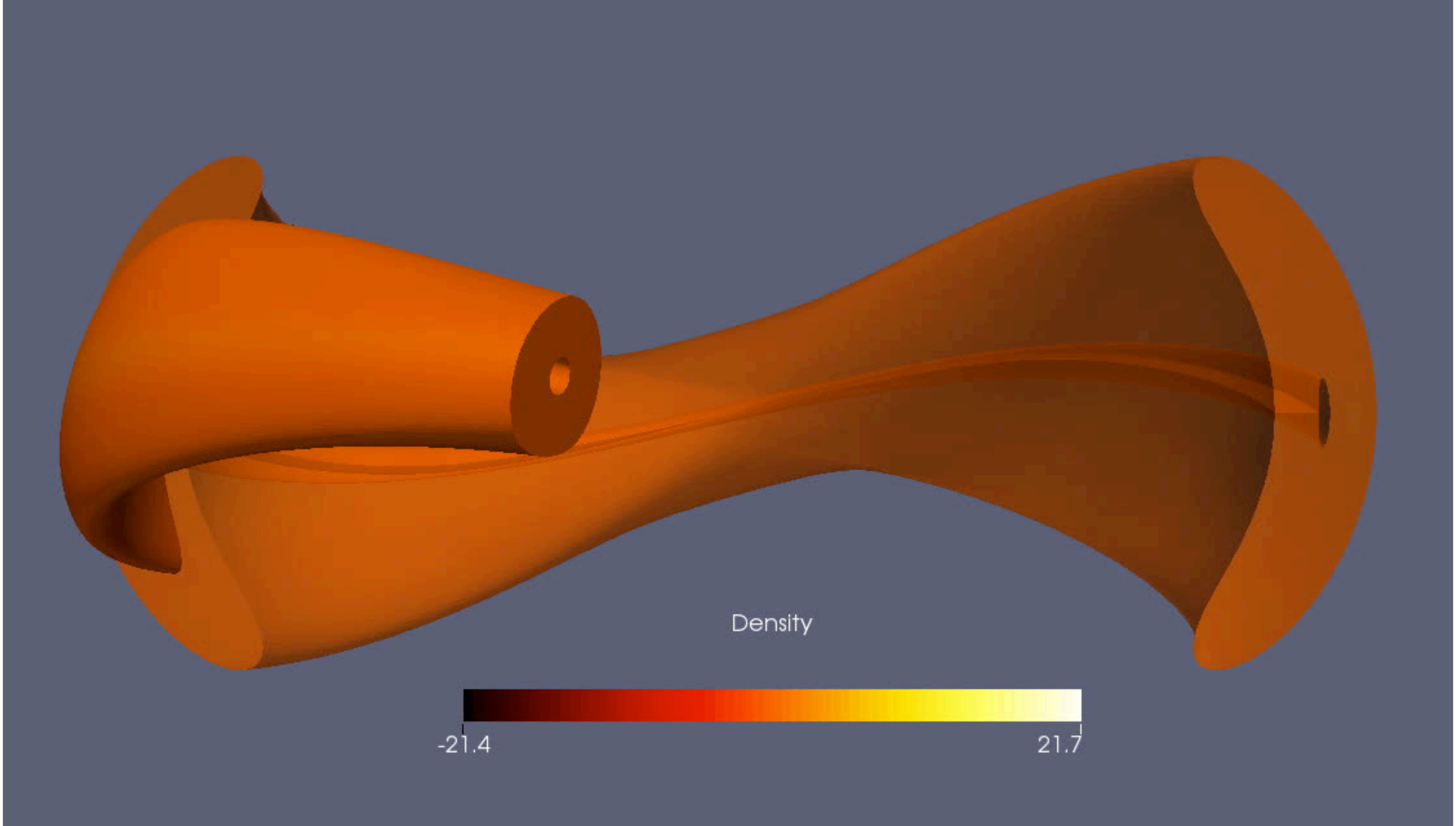
Non-axisymmetric devices: W7-X (2015-)

Complex plasma shape, optimized via HPC

Next step: Take turbulence into account

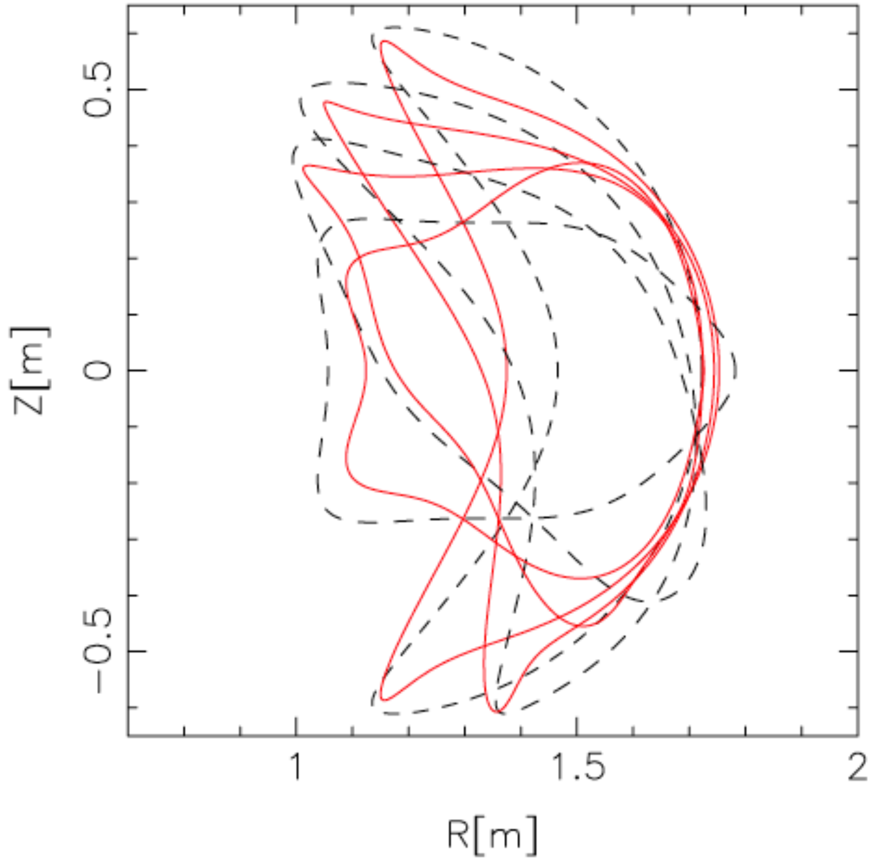


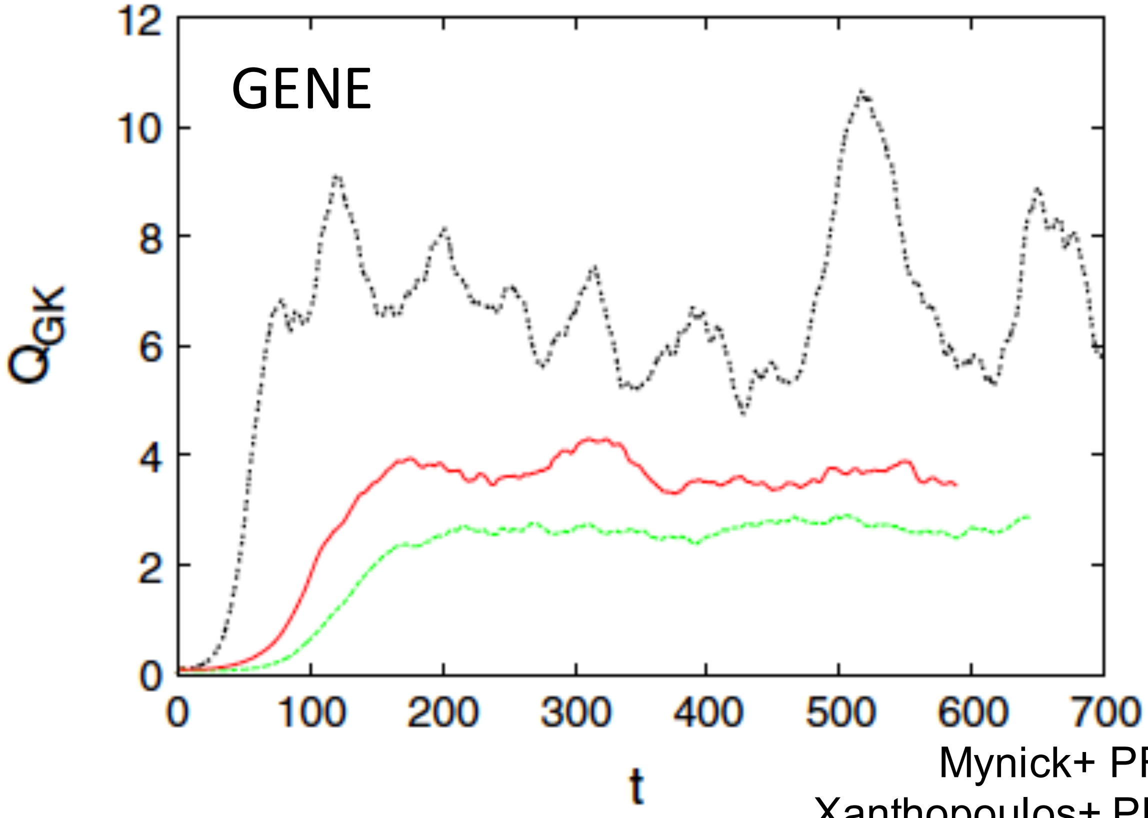



Density
-21.4
21.7

Vision: “Transport-by-design”

Proof-of-principle: Magnetic geometries, optimized for turbulent transport via successive magnetohydrodynamic equilibria, using simple “cost functions” and GENE simulations





Mynick+ PRL 2010
Xanthopoulos+ PRL 2014

Combining computation with data analytics and machine learning for plasma physics

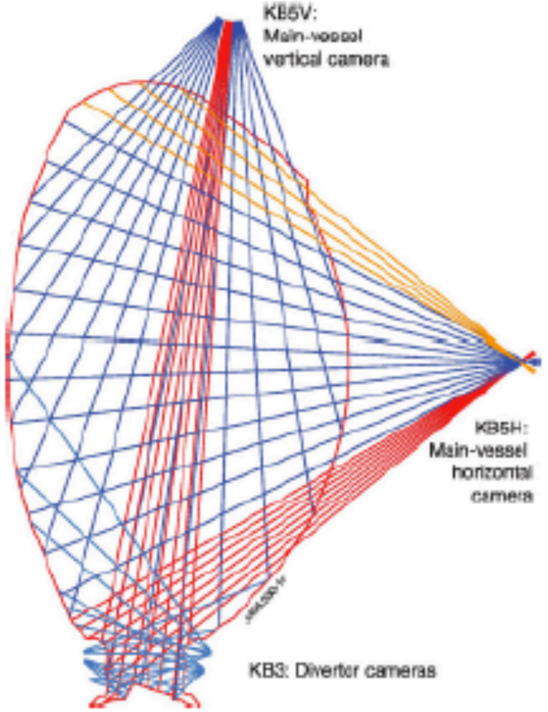
FESAC Report (2018)

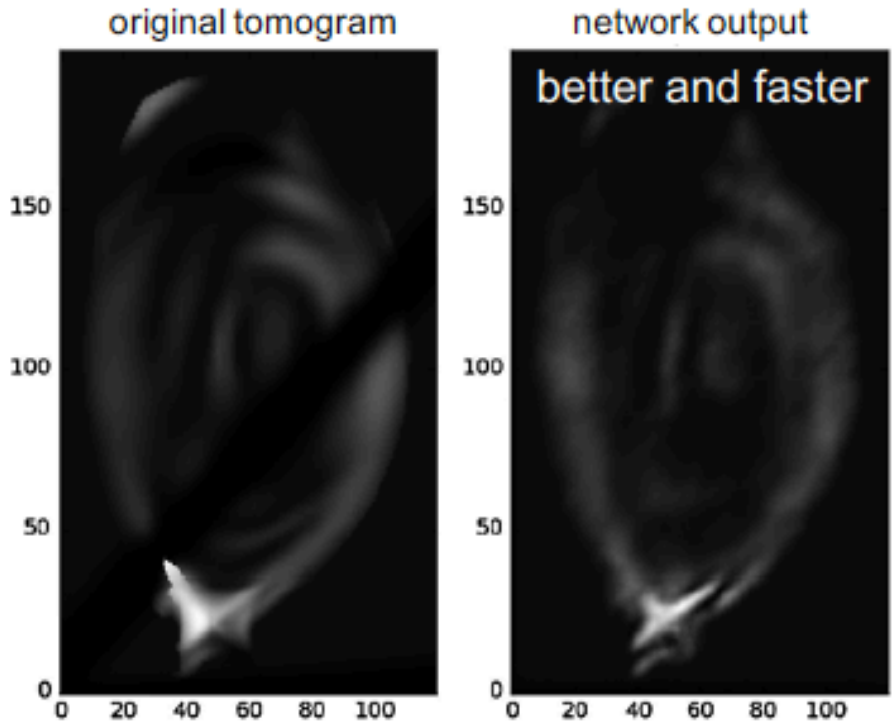
Transformative Enabling Capabilities for fusion

Advanced Algorithms – Advanced algorithms will transform our vision of feedback control for a power-producing fusion reactor. The vision will change from one of basic feasibility to the creation of intelligent systems, and perhaps even enabling operation at optimized operating points whose achievement and sustainment are impossible without high-performance feedback control. The area of advanced algorithms includes the related fields of mathematical control, machine learning, artificial intelligence, integrated data analysis, and other algorithm-based R&D. Given the pace of advances, control solutions that establish fusion reactor operation will become within reach, as will the discovery and refinement of physics principles embedded within the data from present experiments. This TEC offers tools and methods to support and accelerate the pace of physics understanding, leveraging both experimental and theoretical efforts. These tools are synergistic with advances in exascale and other high-performance computing capabilities that will enable improved physics understanding. Machine learning and mathematical control can also help to bridge gaps in knowledge when these exist, for example to enable effective control of fusion plasmas with imperfect understanding of the plasma state.

Deep Learning for real-time plasma control

Plasma tomography: Use CNNs to reconstruct cross-section from projections





Traditional inversion schemes:

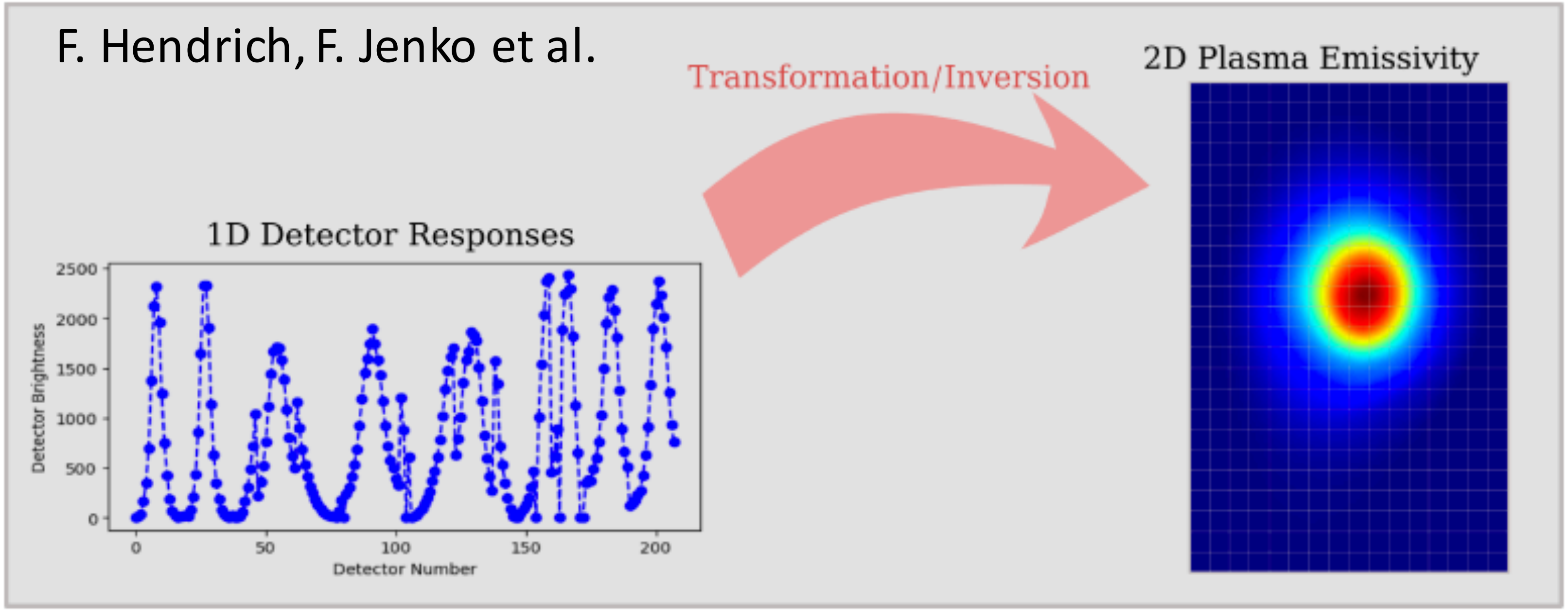
- Varying runtime
- Dependence on additional data (magnetic equilibrium)
- Not real-time capable

Deep learning for plasma tomography using the bolometer system at JET

PhD student at IPP

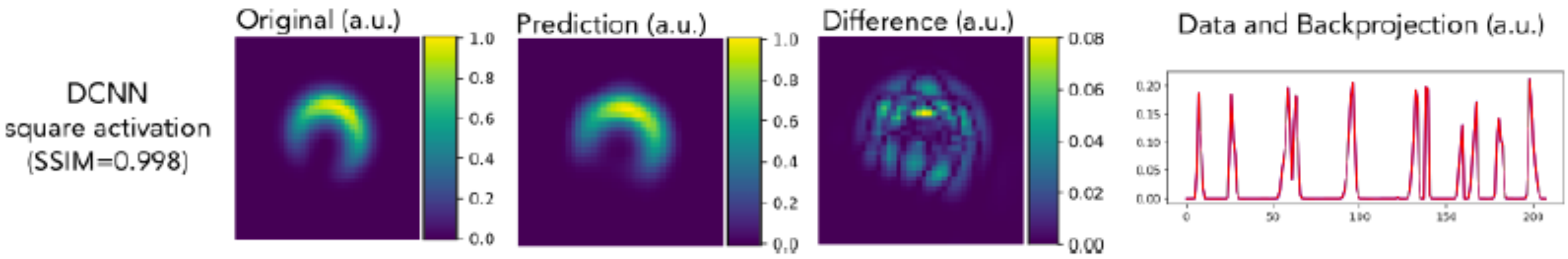
Francisco A. Matos[a], Diogo R. Ferreira[a,*], Pedro J. Carvalho[b], JET Contributors[1]

Unsupervised inversion with Deep Learning



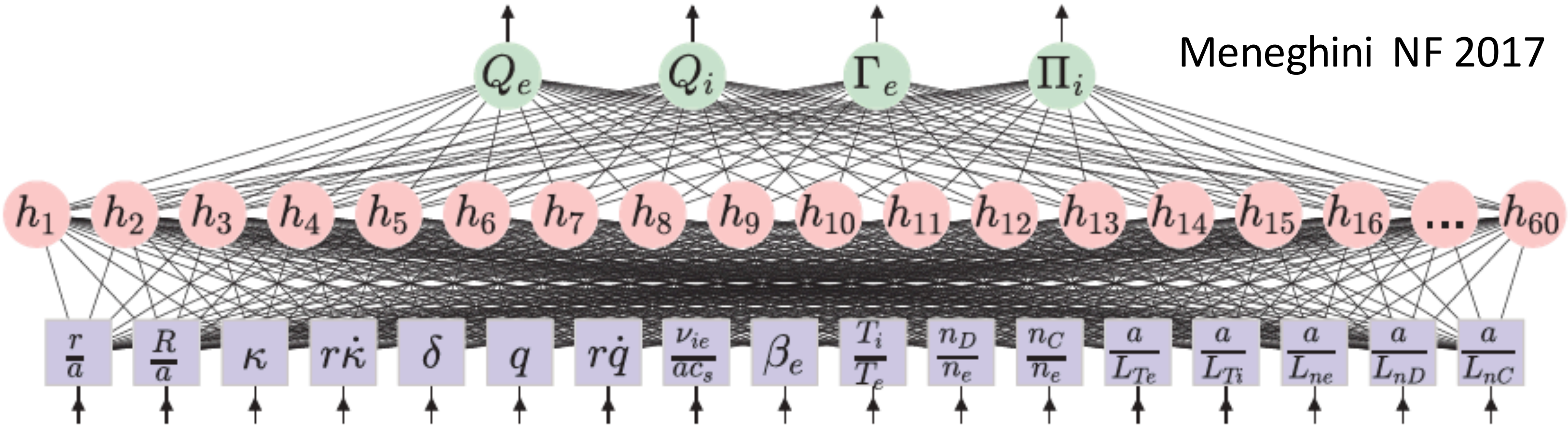
Encode (weak) prior knowledge directly in the NN architecture (in lieu of minimizing an additional regularization term):

- Positivity
- Locality (smoothness)

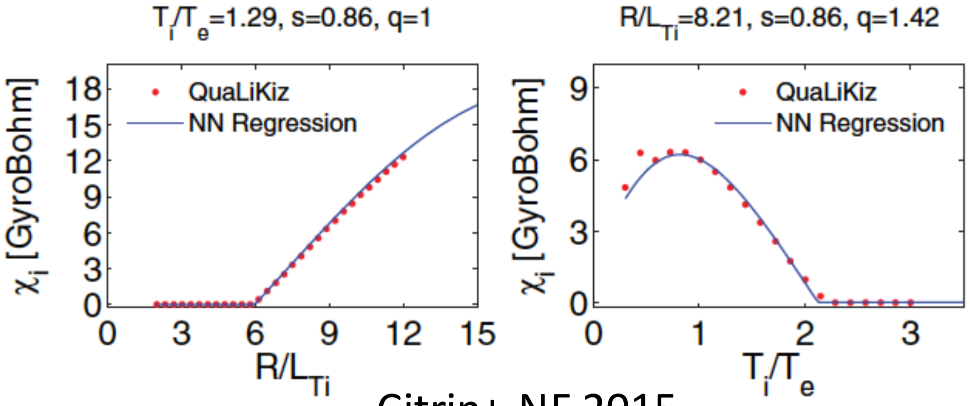


Real-time plasma profile prediction

From *nonlinear* gyrokinetics to *quasilinear* gyrokinetics/gyrofluids to NNs:
Calls for **deep understanding of turbulence** in plasmas



Meneghini NF 2017



Citrin+ NF 2015

NL gyrokinetics: ~10^5 core-h

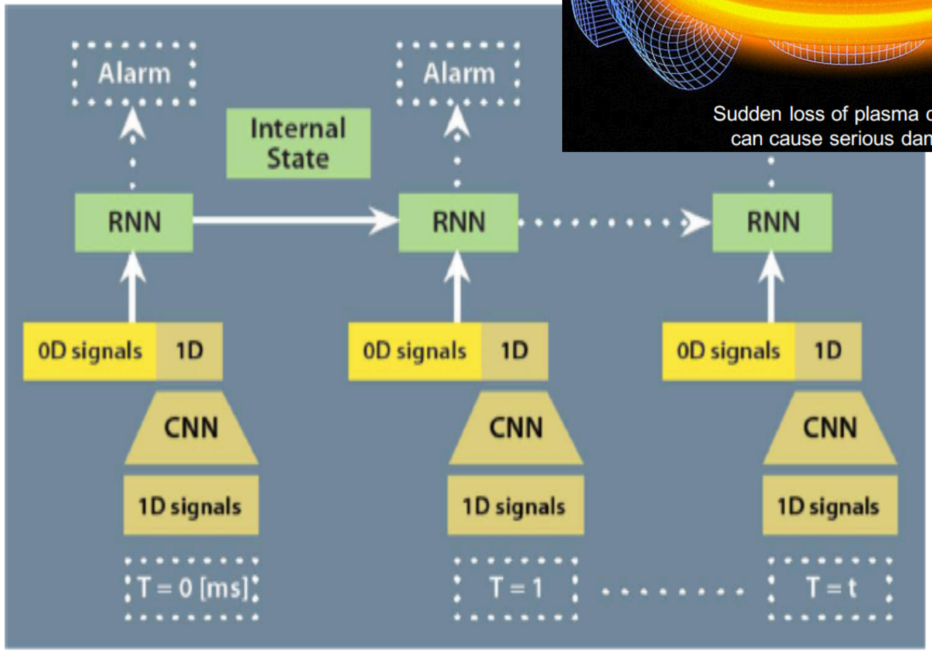
QL gyrokinetics: ~10^{-3} core-h

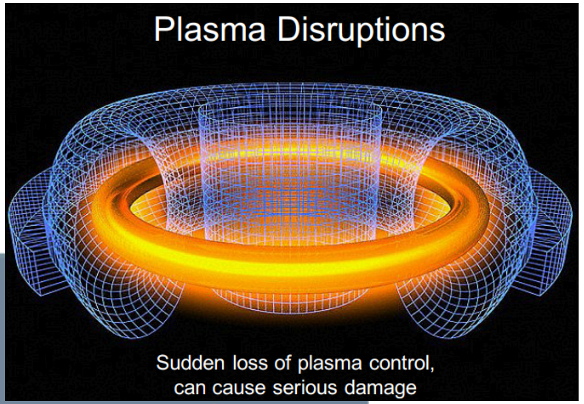
NNs: real-time capability

Similarity to financial data analysis, earthquake prediction etc.

DL for real-time disruption prediction

Recurrent NNs

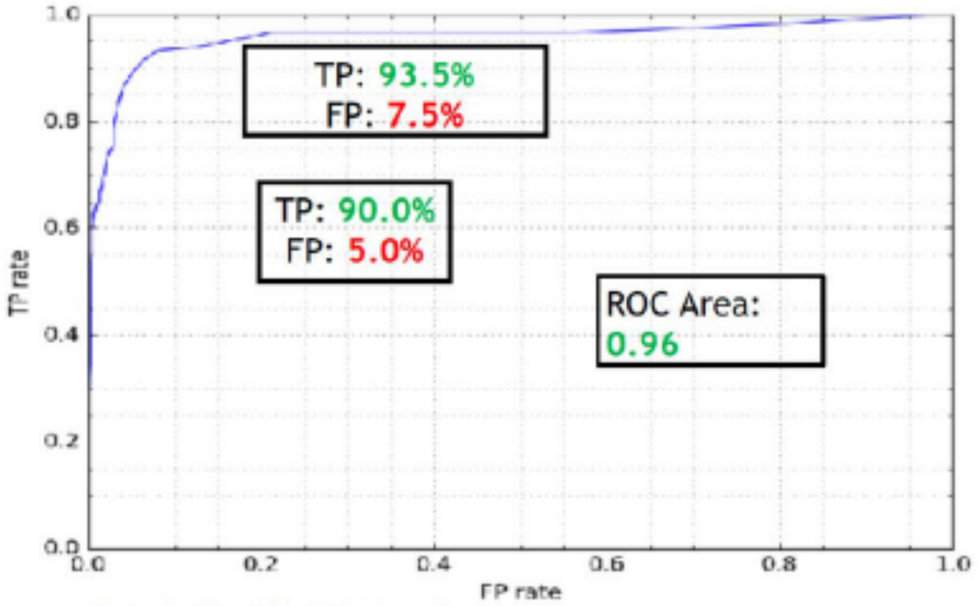




FRNN Code PERFORMANCE: ROC CURVES

JET ITER-like Wall Cases @30ms before Disruption

Performance Tradeoff: Tune **True Positives** (good: correctly caught disruption) **vs. False Positives** (bad: safe shot incorrectly labeled disruptive).



Data (~50 GB), 0D signals:
- Training: on **4100 shots** from **JET C-Wall** campaigns
- Testing **1200 shots** from **Jet ILW** campaigns
- **All shots used**, no signal filtering or removal of shots

This campaign won the NVIDIA Global Impact Award at the 2018 GPU Technology Conference

W. Tang et al. (talk by D. Keyes @ WS1)

Integrated data analysis

R. Fischer

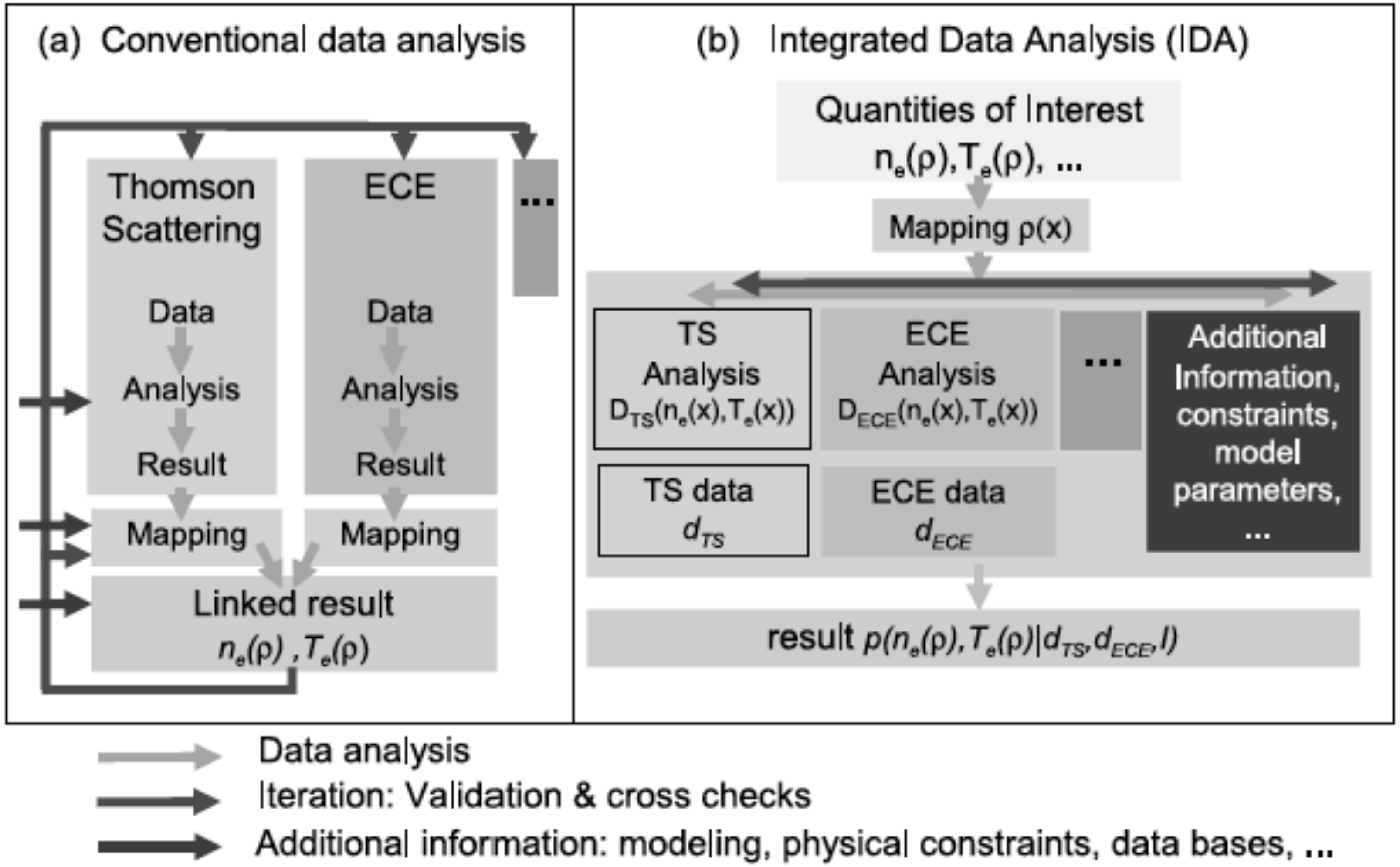
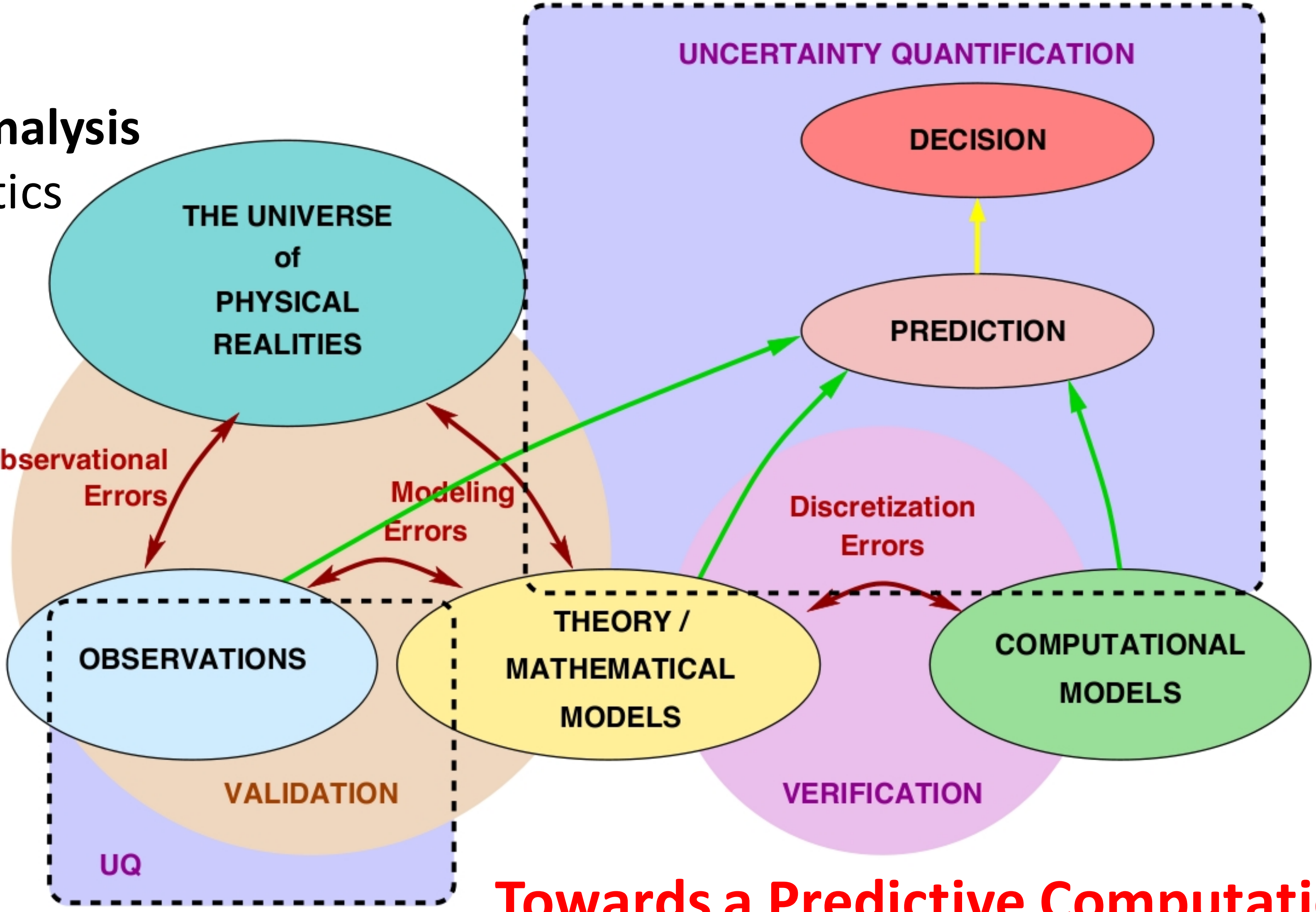


Fig. 1. Schematic comparison of the conventional data analysis with the IDA approach.

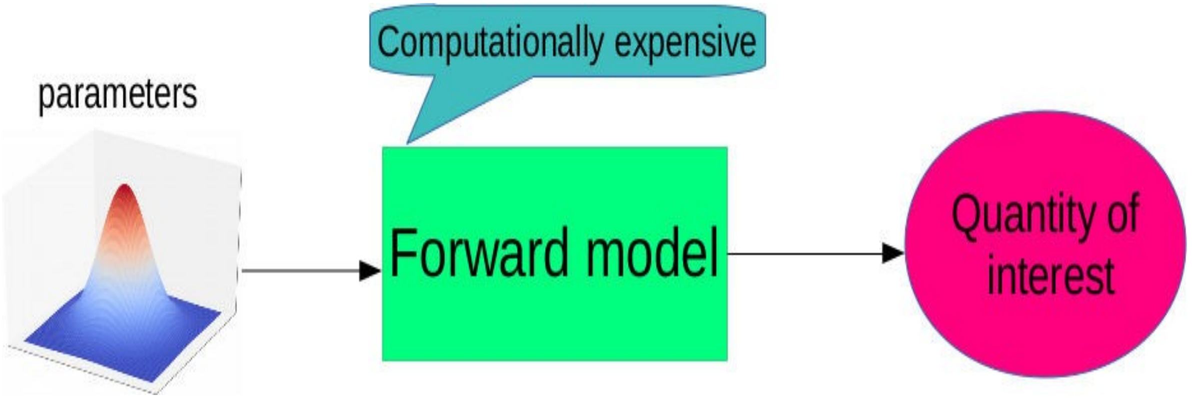
The crucial role of Uncertainty Quantification

Example:
Integrated Data Analysis
via Bayesian statistics



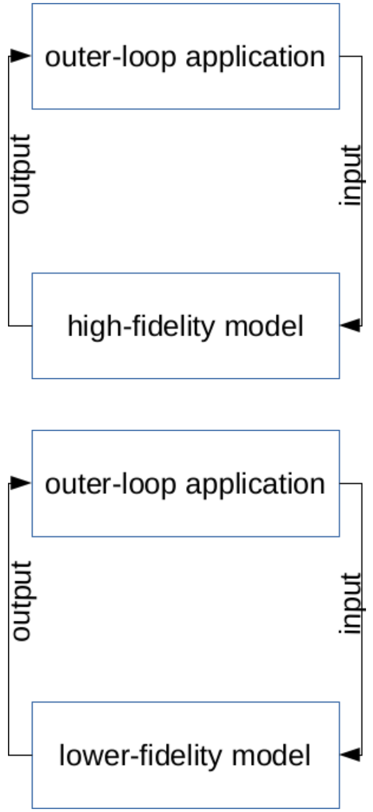
Towards a Predictive Computational Science

Outer-loop applications



Examples: optimization, control, inverse problems, **uncertainty propagation** etc.

Challenges



...usually originate in "high-fidelity" models (e.g., PDEs) having...

— high dimensionality, i.e., large number of stochastic parameters
— high computational demands
— high data throughput

Some solutions

...by replacing the hi-fi model with a "lower-fidelity" model that...

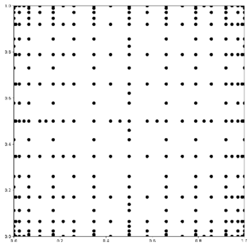
— is less accurate, but **computationally cheap**
— **exploits** the **structure** of the problem at hand

Sensitivity-driven, adaptive sparse approximations

I.-G. Farcas, , T. Görler, H.-J. Bungartz, F. Jenko, T. Neckel, *Sensitivity-driven adaptive sparse stochastic approximations in plasma microinstability analysis*, in preparation

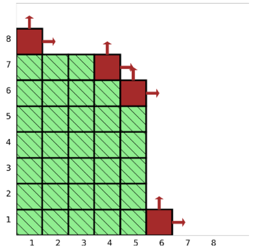
Our approach:

Design a **non-intrusive, sensitivity-driven, adaptive** approach that...



...delays the "curse of dimensionality"

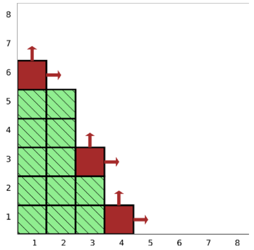
→ **sparse approximations**



...decreases the overall computational cost

→ **sensitivity-driven adaptivity**

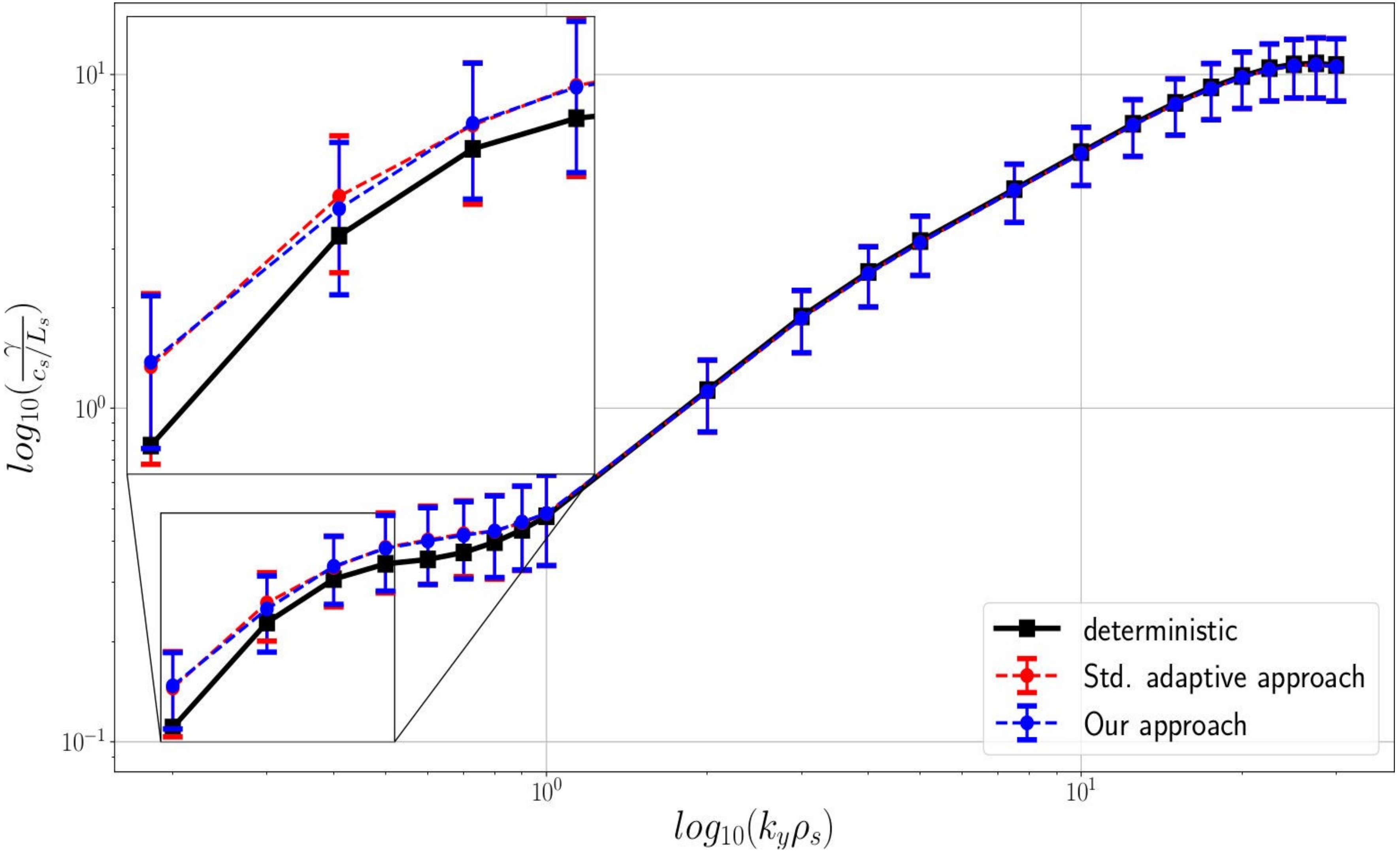
- exploits the structure of the problem at hand
- adds new subspaces based on suitable metrics
- use sensitivity information to drive the adaptive process
- preferentially refine the "important" directions



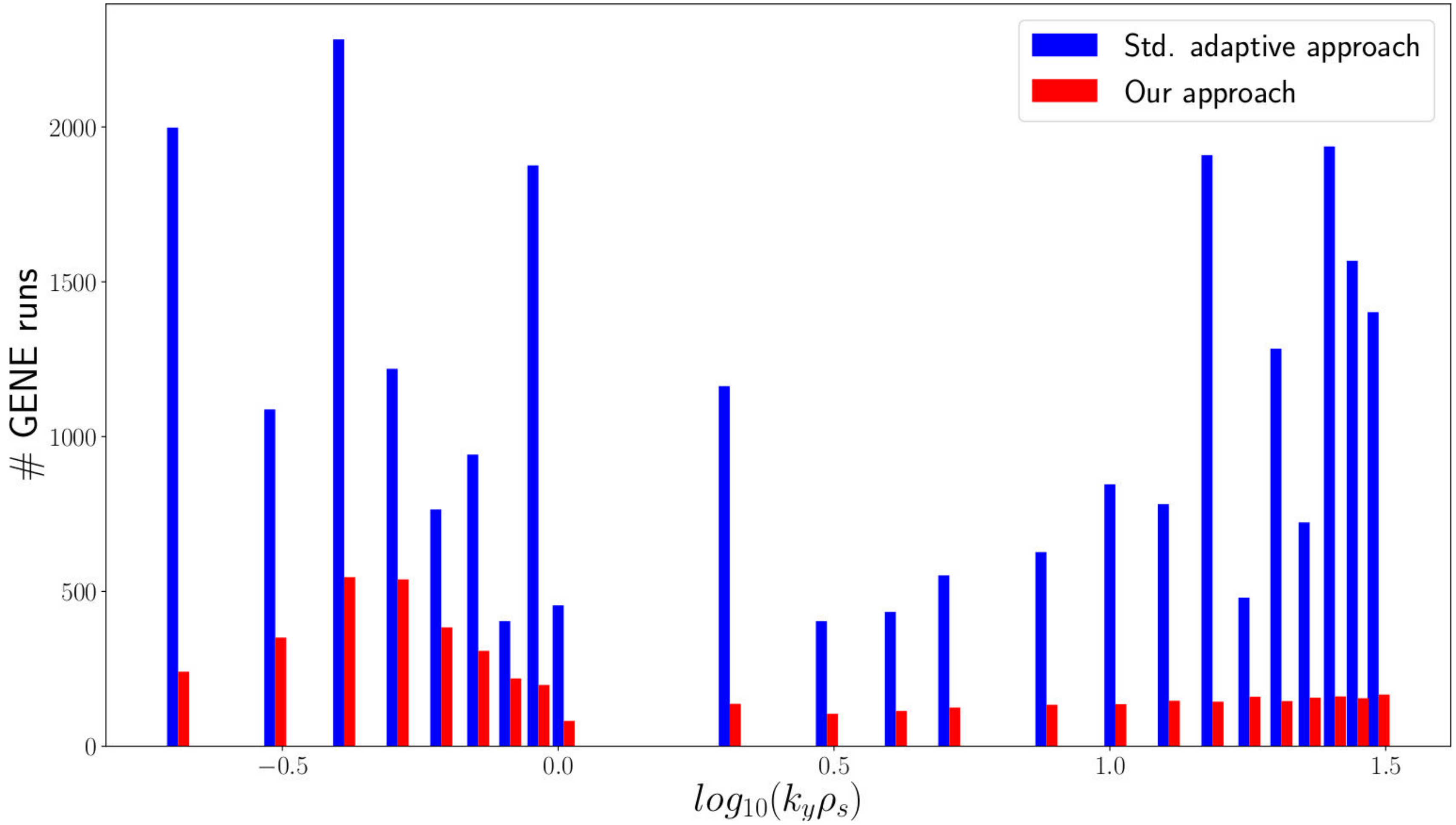
Real-world application: ASDEX Upgrade

stochastic parameter	symbol	left bound	right bound
plasma beta	β	0.488×10^{-3}	0.597×10^{-3}
collision frequency	ν_c	0.641×10^{-2}	0.867×10^{-2}
i/*e* log density gradient	$-(L_s/n)/(dn/dx)$	1.156	1.927
i log temperature gradient	$-(L_s/T_i)/(dT_i/dx)$	2.096	3.494
i temperature	T_i	0.610	0.670
e log temperature gradient	$-(L_s/T_e)/(dT_e/dx)$	4.040	6.733
effective ion charge	$Z_{eff} = \sum_i n_i q_i^2 / \sum_i n_i$	1.280	1.920
safety factor	q	2.170	2.399
magnetic shear	$\hat{s} = \frac{r}{q}\frac{dq}{dr}$	1.992	2.435
elongation	k	0.128	0.141
elongation gradient	$s_k = \frac{r}{k}\frac{\partial k}{\partial r}$	0.200	0.250
triangularity	δ	0.710	0.870

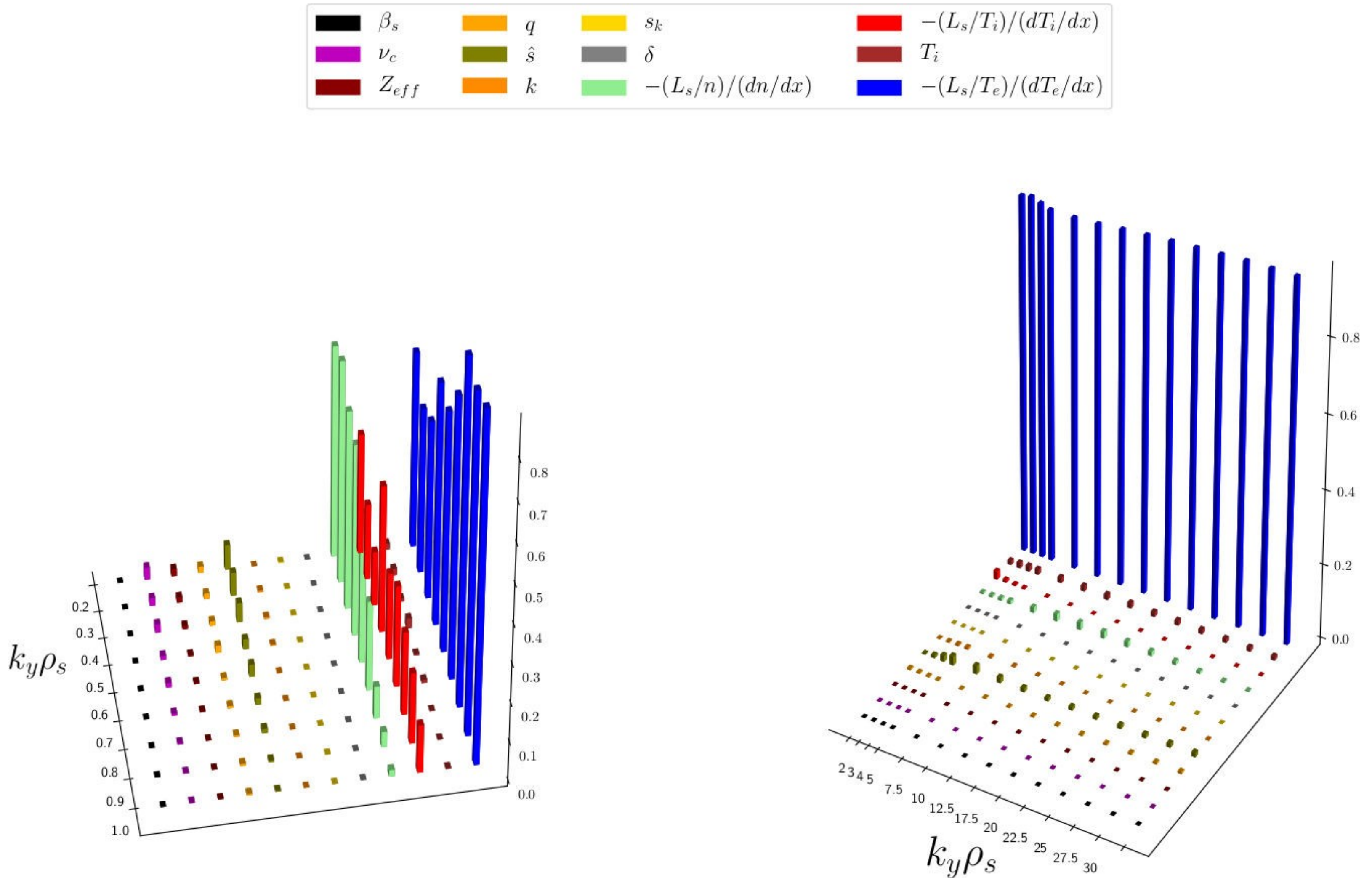
Linear growth rates of microinstabilities (GENE)



Number of necessary GENE runs is minimized

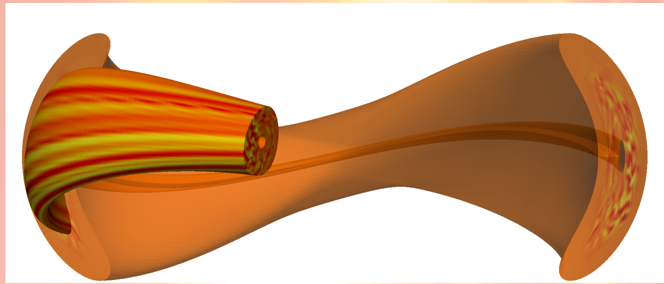


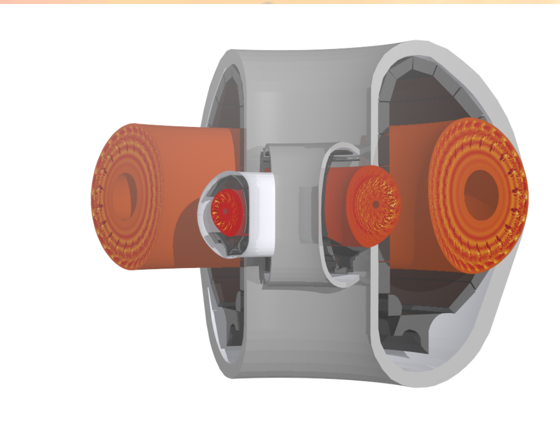
Variable importance: total Sobol' indices



Summary

The big picture





Overarching goal: Contribute to the **gradual development of a validated predictive capability ("virtual fusion plasma"), helping to accelerate fusion energy research**

A beautiful example of how **fascinating science on some of the world's largest supercomputers contributes to solving grand challenges facing society**

Growing role of (Theory-Guided) Data Science; towards a convergence with HPC?