

GAMM 2026 @ STUTTGART

Data-Driven Spectral Prediction for Accelerating Large-Scale Electronic Structure Calculations

November 10, 2026 | Gustavo Ramirez-Hidalgo | Jülich Supercomputing Center, JSC

Member of the Helmholtz Association

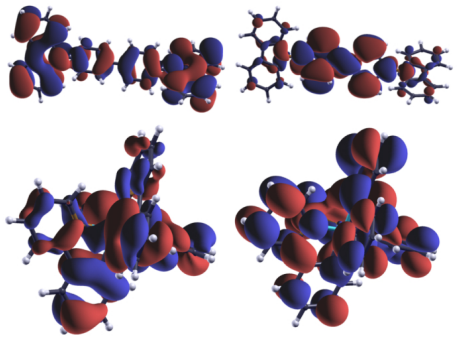
JÜLICH Forschungszentrum

EuroHPC Joint Undertaking

INNO4SCALE

THE LIMITX CONSORTIUM

- Germany - Jülich Supercomputing Centre (JSC): Edoardo Di Napoli, Xinzhe Wu, Gustavo Ramirez-Hidalgo.
- Croatia - Ruder Boskovic Institute (RBI): Davor Davidovic, Abhiram Badrinarayanan, Jurica Novak.
- France - CEA / IRIG / MEM / L_Sim: Luigi Genovese.



JÜLICH Forschungszentrum EuroHPC Joint Undertaking INNO4SCALE

THE COMPUTATIONAL BOTTLENECK

The Challenge: Simulating large molecular systems with (tens/hundreds of) thousands of atoms requires highly scalable methodologies.

- **Density Functional Theory (DFT):** Codes like BigDFT scale linearly, but major hurdles remain.
- **The Bottleneck:** Solving large, sparse linear systems and complex eigenproblems: $Hv_\lambda = \lambda S v_\lambda$.
- **Traditional Approaches:** Diagonalization-based methods struggle with the sheer scale of exascale computing demands.
- **Our Goal:** Develop a robust predictor of spectral properties, and ultimately optimize the eigensolver.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

THE LIMITX DATA STRATEGY

Dataset Metrics

- **660** Protein Dimer Computations
- **∼2 TB** Generated Data
- **15,000+** Max Atoms per System

Fragment-Based Representation
To efficiently train ML/DL models on massive molecular systems, we moved away from computationally prohibitive all-atom representations.

Instead, we abstract the physical systems into **amino acid-based fragments**. Our models utilize adjacency matrices that encode inter-fragment distances, **drastically reducing data complexity** while preserving essential structural relationships.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

OVERCOMING DIMENSIONALITY

The Challenge

Predicting the full eigenspectrum of a system with thousands of individual eigenvalues is computationally infeasible. Neural networks struggle to learn when the output vector length varies drastically depending on the size of the input molecular system.

The Innovation

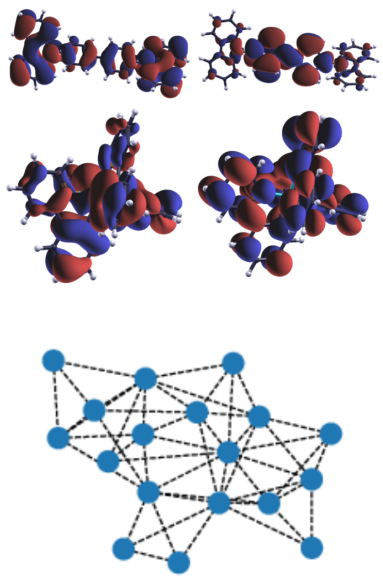
Instead, our models are trained to predict the coefficients of a **Chebyshev polynomial** (that interpolates the eigenspectrum).
This transforms the problem into predicting a **fixed-length vector** of coefficients, ensuring a scalable, efficient, and generalizable learning process regardless of system size.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

SHORT RECAP: INPUTS AND OUTPUTS

We are shrinking data, both at the input and output levels, to make the processes of training and predicting much more tractable.

Input



Output

Eigenvalues
(possibly (tens/hundreds of) thousands)

$\downarrow$

$\lesssim 100$ Chebyshev coefficients

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

MACHINE LEARNING MODELS EXPLORED

We investigated three distinct predictive approaches:

- **Kernel Ridge Regression (KRR - the cheapest):**
  An interpretable model utilizing randomized compression on the spectrum of Coulomb matrices as input. Optimized via comprehensive grid search.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

MACHINE LEARNING MODELS EXPLORED

We investigated three distinct predictive approaches:

- **Kernel Ridge Regression (KRR - the cheapest):**
  An interpretable model utilizing randomized compression on the spectrum of Coulomb matrices as input. Optimized via comprehensive grid search.
- **Graph Neural Networks (GNN - the most expensive):**
  Based on a modified SchNet architecture. Nodes represent amino acid types, and edges encode inter-fragment distances. Proved highly effective at predicting the overall shape.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

MACHINE LEARNING MODELS EXPLORED

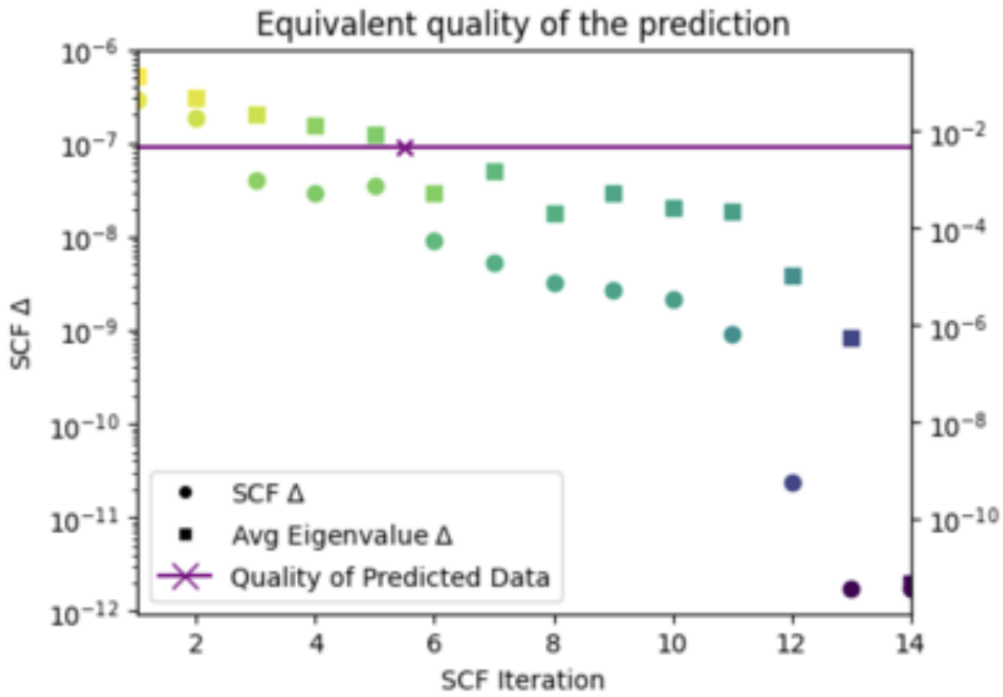
We investigated three distinct predictive approaches:

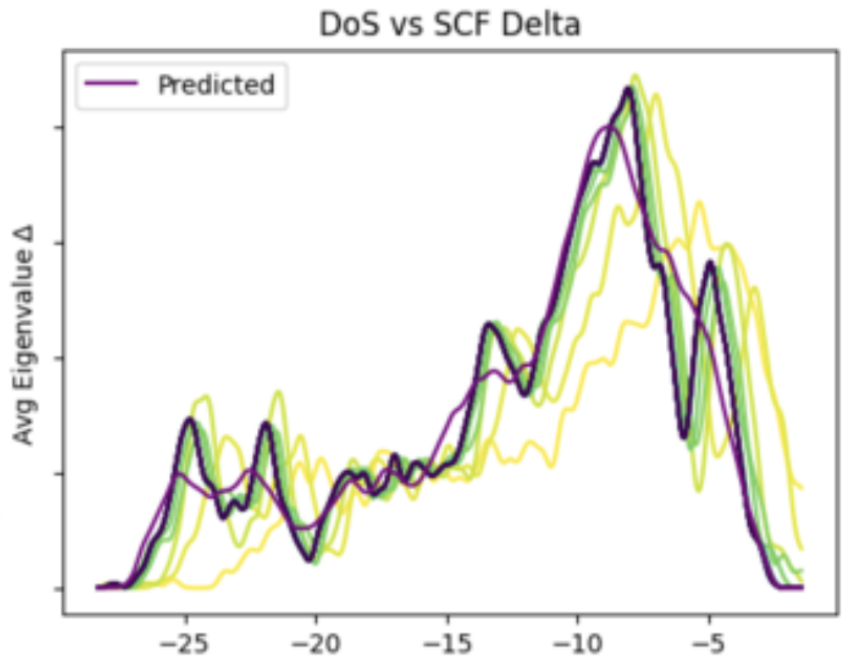
- **Kernel Ridge Regression (KRR - the cheapest):**
  An interpretable model utilizing randomized compression on the spectrum of Coulomb matrices as input. Optimized via comprehensive grid search.
- **Graph Neural Networks (GNN - the most expensive):**
  Based on a modified SchNet architecture. Nodes represent amino acid types, and edges encode inter-fragment distances. Proved highly effective at predicting the overall shape.
- **RDF + Random Forest:**
  A fast approach utilizing a weighted Radial Distribution Function (RDF). Incorporates electronic properties like electronegativity directly into the structural descriptor.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

SOME RESULTS

LimitX system 1ZSG (1861 atoms).





JÜLICH Forschungszentrum

EuroHPC Joint Undertaking

INNO4 SCALE

INTERPLAY WITH EIGENSOLVING

Enhancing rational filter-based eigensolvers via data-driven strategies:

- **Precise Subspace Allocation via DOS Prediction:**
  The various models accurately estimate the DoS. This allows FrASE to precisely allocate subspace sizes.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

INTERPLAY WITH EIGENSOLVING

Enhancing rational filter-based eigensolvers via data-driven strategies:

- **Precise Subspace Allocation via DOS Prediction:**
  The various models accurately estimate the DoS. This allows FrASE to precisely allocate subspace sizes.
- **Krylov Subspace Reuse & Polynomial Preconditioning:**
  Using ML-predicted spectral data, we can construct robust polynomial preconditioners, accelerating linear solves across multiple shifted systems simultaneously without costly factorizations.

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

INTERPLAY WITH EIGENSOLVING

Enhancing rational filter-based eigensolvers via data-driven strategies:

- **Precise Subspace Allocation via DOS Prediction:**
  The various models accurately estimate the DoS. This allows FrASE to precisely allocate subspace sizes.
- **Krylov Subspace Reuse & Polynomial Preconditioning:**
  Using ML-predicted spectral data, we can construct robust polynomial preconditioners, accelerating linear solves across multiple shifted systems simultaneously without costly factorizations.
- **Adaptive Solver Selection:**
  Predicted spectral insights guide dynamic algorithm choices (e.g., dense regions trigger block iterative solvers with heavy preconditioners, sparse regions employ lightweight solvers, etc.).

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE

Thank you!
Questions?

https://www.inno4scale.eu/success-story-limitx/

g.ramirez.hidalgo@fz-juelich.de

JÜLICH Forschungszentrum
EuroHPC Joint Undertaking
INNO4 SCALE