NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Examples of Machine Learning to Process and Model Large Datasets

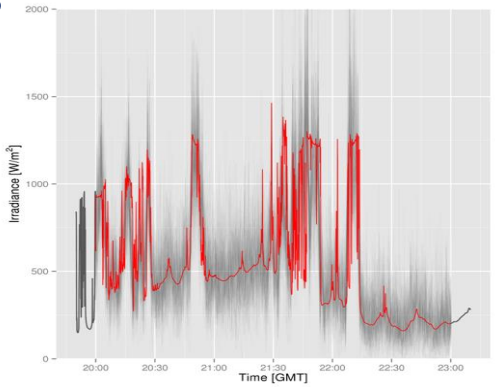
Big Data and Machine Learning for Clean Coal and Carbon Management Strategic Planning Workshop

July 12, 2018

Zia Abdullah
Bioenergy Laboratory Program Manager
NREL

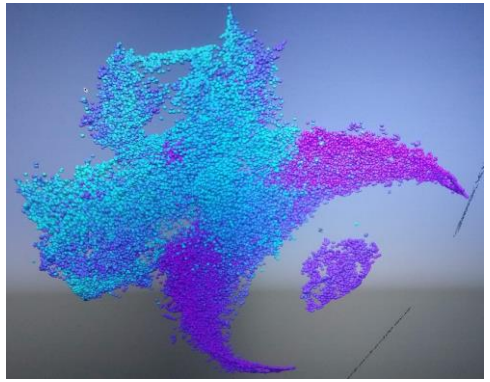
Machine-learning and statistical techniques enable cleansing, discovery, and modeling of massive datasets.

- Anomaly-detection and imputation methods automate the cleansing of large datasets.



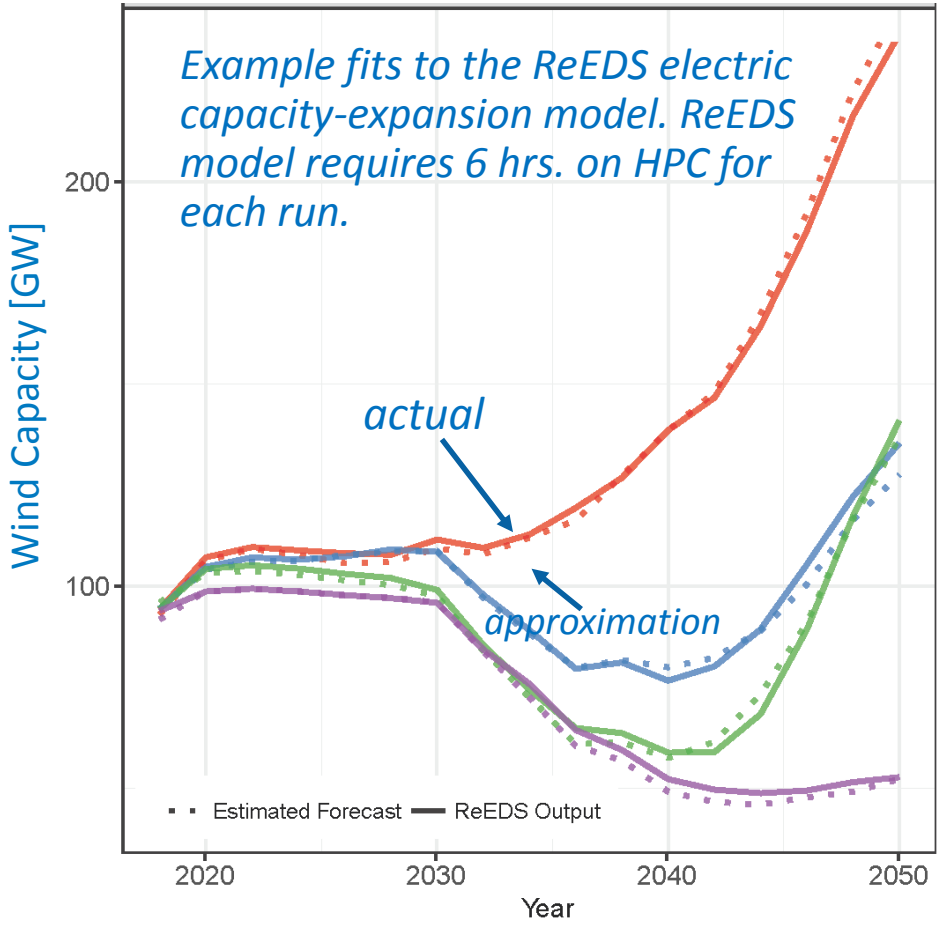
Reconstruction (black) of missing solar irradiance data (red) using a wavelet-based multi-scale covariance model.

- Dimension-reduction techniques and self-organized maps identify insightful patterns and low-dimensional features embedded in big datasets.



Abstract 3D feature visualization of a 50-dimensional dataset of 1.8 million simulations of bioenergy industry scenarios.

- Deep neural networks provide fast, high-quality approximations to state-of-the-art energy simulations.



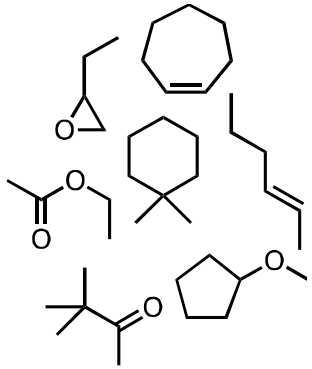
ReEDS: Regional Economic Development Scenario

Big Data is used to Train ANNs to Predict Material Properties from Molecular Structure

Accelerate the Design of performance-advantaged biomass-derived polymers

Explain novel behavior of bio-derived polymers seen in laboratory

Predict and Suggest new formulations for bio-derived polymers to synthesize in lab

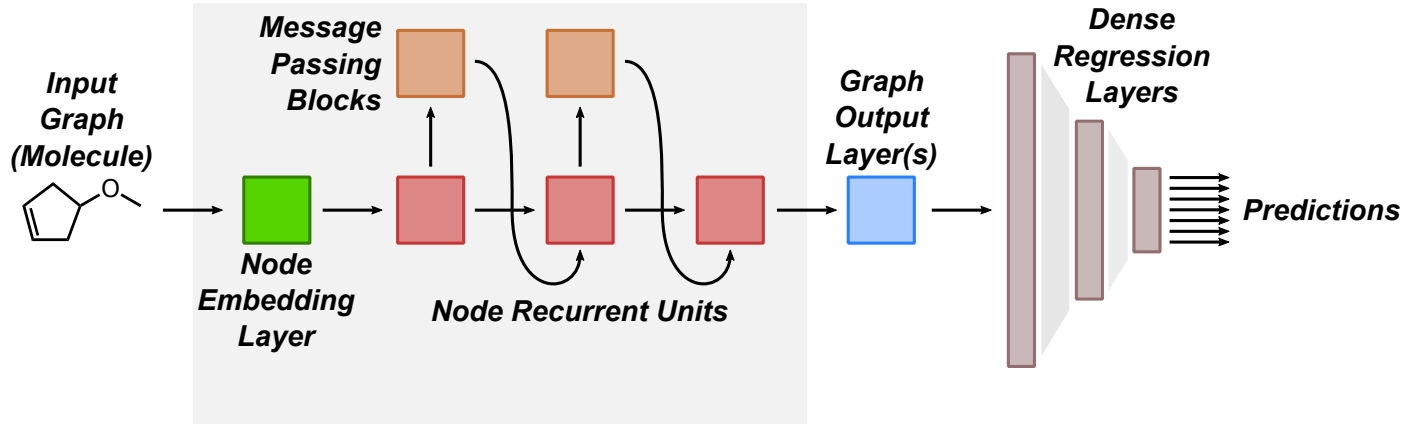


Traditional Quantitative Structure Property Relationship learning

Molecular Descriptors / Fingerprints → $\begin{pmatrix} u_1 \\ v_1 \\ w_1 \\ \vdots \\ u_n \\ v_n \\ w_n \end{pmatrix}$ → Traditional Machine Learning

End to end learning → **Property Prediction**

Single step from atomic structure to prediction.
Optimal molecular representation 'learned' from the data



Data
- Chemistry Molecular Databases
- Polymer Property Databases
- Molecular Dynamics and QM
- Experiments

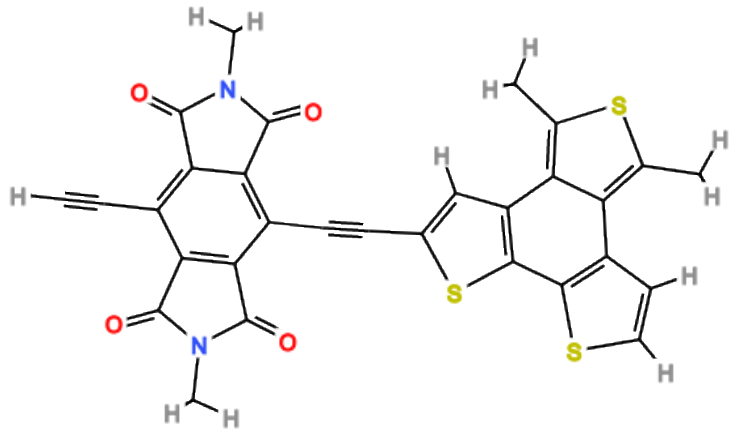
Machine Learning
- Transfer Learning from any Chemical Database

Application of ANNs to Predict Orbital Energies in Organic Photovoltaic Solar Cells

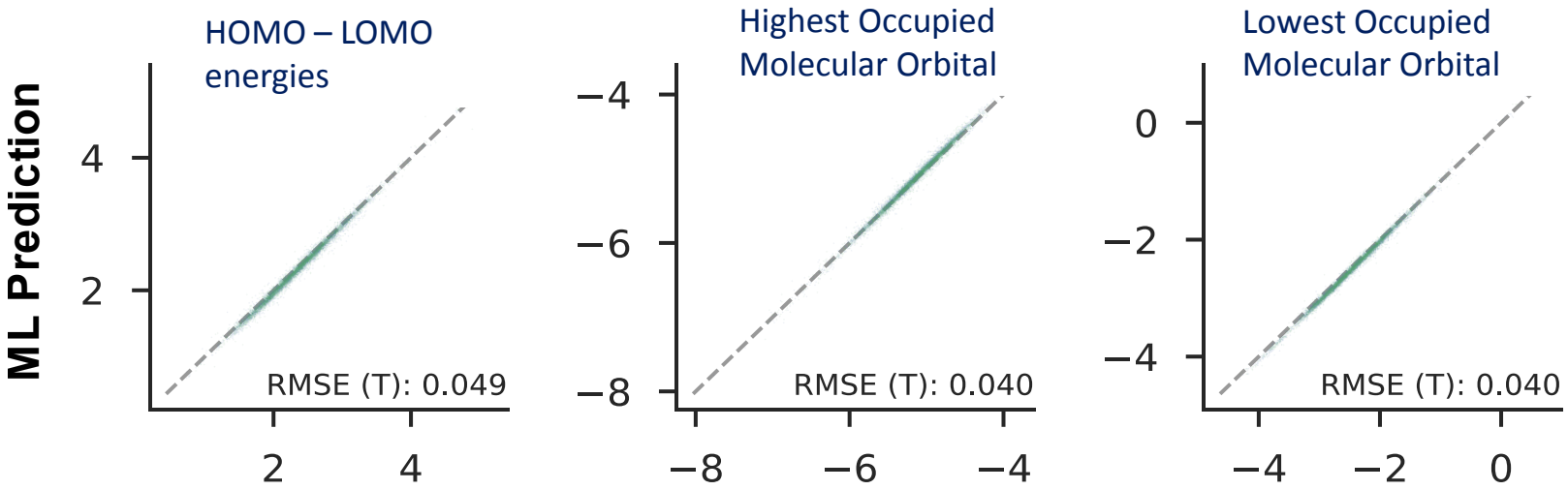
NREL NATIONAL RENEWABLE ENERGY LABORATORY

Powered by SunShot U.S. Department of Energy

Computational Database for Active Layer Materials for Organic Photovoltaic Solar Cells



>90,000 monomers,
>50,000 with extrapolated polymer results



- Orbital energies predicted to ~1 kcal/mol (approximately experimental error)
- Computation times reduced by ~6 orders of magnitude: 10^{-3} s for ML, 10^{3} s for DFT

DFT, Density Functional Theory', is the numerically intensive computational result which is predicted with machine learning.

units in eV

www.nrel.gov

NREL
NATIONAL RENEWABLE ENERGY LABORATORY