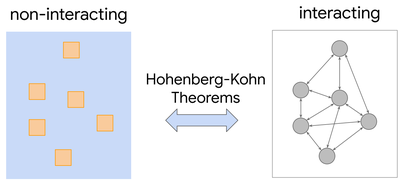
Kohn-Sham equations as regularizer: building prior knowledge into machine-learned physics

Li Li (李力)[1]*, Stephan Hoyer[1], Ryan Pederson[2], Ruoxi Sun[1], Ekin D. Cubuk[1], Patrick Riley[1], Kieron Burke[2]
[1]Google Research, Mountain View, CA 94043, USA [2]University of California, Irvine, CA 92697, USA * leeley@google.com

Density Functional Theory

Density functional theory (DFT), an approach to electronic structure problems, took an enormous step forward with the creation of the Kohn-Sham (KS) equations, which greatly improves accuracy from the original DFT. The results of solving the KS equations are reported in tens of thousands of papers each year.

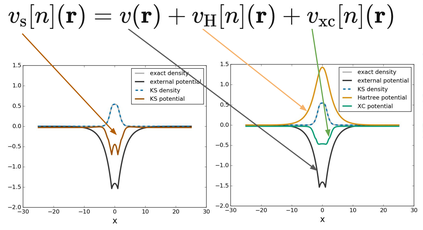


$$E[n] = T_s[n] + V[n] + U[n] + E_{XC}[n]$$

Kohn-Sham equation

$$\left\{-\frac{\nabla^2}{2} + v_s[n](\mathbf{r})\right\}\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r})$$

$$v_s[n](\mathbf{r}) = v(\mathbf{r}) + v_H[n](\mathbf{r}) + v_{xc}[n](\mathbf{r})$$



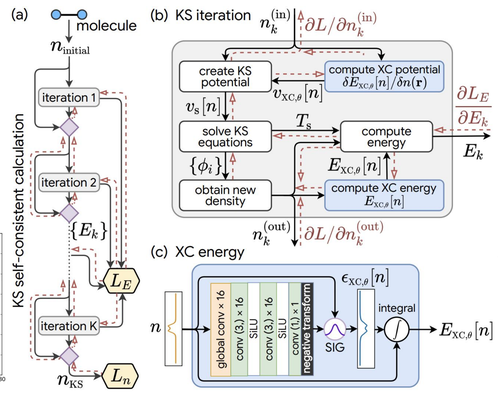
Exchange-correlation (XC) energy functionals

- The only approximation in DFT
- Jacob's ladder:
  - Local density approximation (LDA)
  - Generalized gradient approximation (GGA)
  - meta-GGA, hybrid, ...
- Over 200 forms of XC exist!

K. Burke. JCP (2012).

Impact of DFT

- The Nobel Prize in Chemistry 1998
- Nvidia estimates that *15%* of world's supercomputer time is devoted to DFT calculations.
- Over *40k* papers published per year.
- The algorithm behind many dataset for machine learning models: QM9, Open Catalyst Project, Materials Project...



KS self-consistent calculation as a differentiable program

- Differentiate the entire calculation end-to-end using JAX-DFT.
- Self-consistency <-> RNN
- Density mixing <-> residual connection
- Intentions behind loss function:
  - KS self-consistent calculations would output the exact density;
  - the intermediate energies over iterations would converge to the exact energy.

$$L(\theta) = \underbrace{\mathbb{E}_{\text{train}}\left[\int dx (n_{KS} - n_{DMRG})^2/N_e\right]}_{\text{density loss } L_n} + \underbrace{\mathbb{E}_{\text{train}}\left[\sum_{k=1}^{K} w_k (E_k - E_{DMRG})^2/N_e\right]}_{\text{energy loss } L_E}$$

Learning Functionals on One-dimensional H_2

Not an easy task

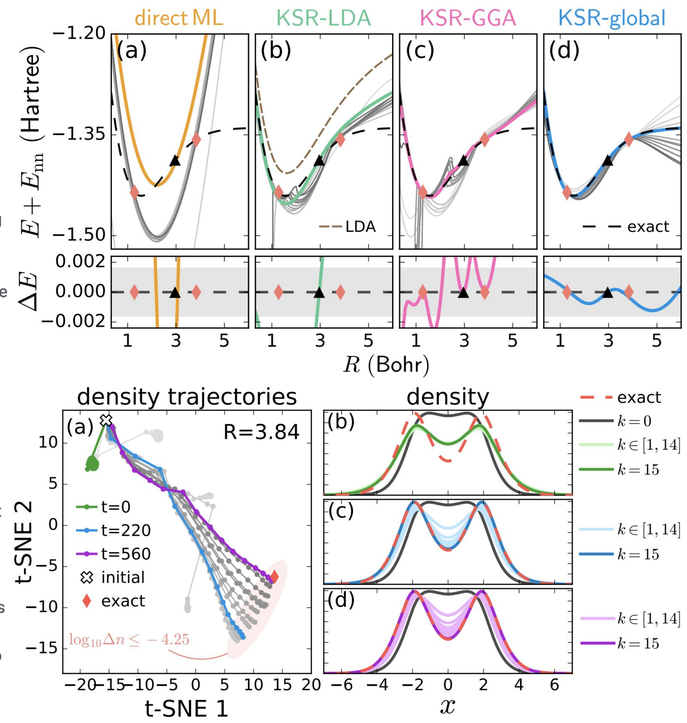
- The regions around equilibrium and the stretched limit are dominated by two different physics.
- The stretched H2 are strongly correlated system, where most density functionals fail.

Performance

- Direct ML clearly fails to capture the physics from very limited data.
- KSR-LDA and -GGA yield reasonably accurate results near the training data, but not in the stretched limit.
- KSR-global achieves chemical accuracy for all separations. The distribution of the prediction at different checkpoints (grey curves) are near the true dissociation curve.

Why Kohn-Sham regularizer (KSR) works with limited data?

- Neural networks are usually underdetermined so regularization is crucial to improve generalization especially when data is limited.
- The physics computation procedure itself can act as a regularizer by sampling the density space during the training.
- Most existing ML functionals learn to predict a single iteration from the exact density, which is a poor surrogate for the full self-consistent calculations.
  - The model is not expected to behave correctly on unseen initial densities for KS calculations.
  - A model trained on many densities for single iteration prediction is not guaranteed to converge in self-consistent calculations..
- Since KSR allows the model access to all the KS iterations, it learns to optimize the entire self-consistent procedure to avoid the error accumulation from greedy optimization of single iterations.



Neural XC Functionals

Conventional XC functionals

symbolic form from physics intuition + some (or no) fitting parameters

Neural XC functionals

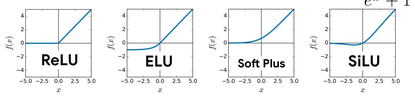
differentiable components from physics intuition + some (or many) trainable parameters

Differentiable components with physics intuition

- Global convolution capturing the long range interaction

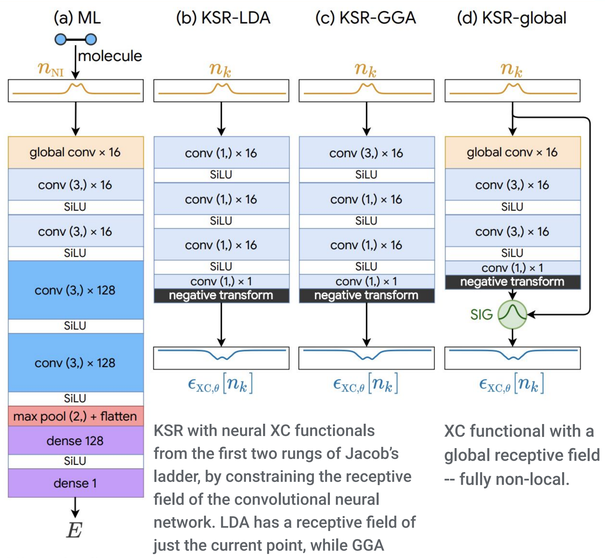
$$G(n(x), \xi_p) = \frac{1}{2\xi_p}\int dx' n(x') \exp(-|x - x'|/\xi_p)$$

  - Two special cases $G(n(x), \kappa^{-1}) \propto \epsilon_H$, $G(n(x), 0) = n(x)$
- Activation function $f(x) = \frac{xe^x}{e^x + 1}$



- Gauge on ϵ_{XC} :
  - vanishing as density goes to zero; negativity
- Self-interaction gate (SIG)

$$\epsilon_{XC}^{(out)} = \epsilon_{XC}^{(in)}(1 - \beta) - \epsilon_H\beta$$
$$\beta(N_e) = \exp(-(N_e - 1)^2/\sigma^2)$$



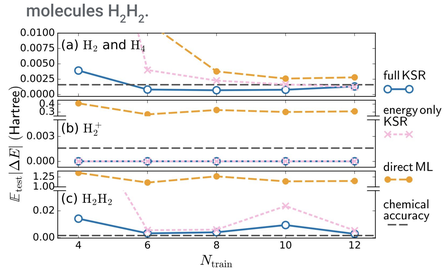
ML model directly predicting E from geometries.

KSR with neural XC functionals from the first two rungs of Jacob's ladder, by constraining the receptive field of the convolutional neural network. LDA has a receptive field of just the current point, while GGA includes the nearest neighbor points, the minimal information for computing the spatial gradient of the density.

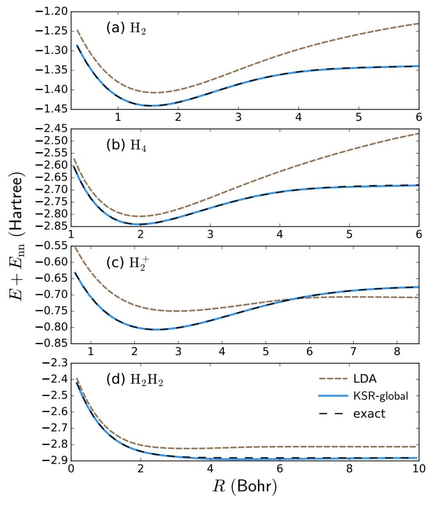
XC functional with a global receptive field -- fully non-local.

Generalization to Unseen Molecules and Outlook

- Retrain neural XC functional with KSR on N_{train} = 8 examples each of H2 and H4 molecules.
  - full KSR: KSR with both energy and density loss
  - energy only KSR: KSR with only energy loss
  - direct ML
- Full KSR has the lowest error at minimum N_{train} = 4, reaching chemical accuracy at 6.
- As the size of the training set increases, energy only KSR reaches chemical accuracy at 10, but direct ML model never does (even at 20).
- Both KSR models have perfect prediction on H_2^+ because of the self-interaction gate while direct ML models always have large errors.
- KSR models generalize much better than ML on unseen molecules H_2H_2.



Dissociation curves from the neural XC functional trained from N_{train} = 8



Outlook

- Our results for KS-DFT serve as proof of principle for rethinking computational physics in this new era with enormous development in
  - automatic differentiation libraries (JAX, PyTorch, TensorFlow),
  - hardware accelerators (GPUs, TPUs),
  - and deep learning algorithms.
- Physics computation itself serves as an implicit regularizer while training the embedded neural networks.
- Differentiable programming blurs the boundary between physics computation and ML. Beyond density functionals, in principle all heuristics in DFT calculations, e.g., initial guess, density update, preconditioning, basis sets, even the entire self-consistent calculations as a meta-optimization problem.
- Code and examples: github.com/google-research/google-research/tree/master/jax_dft