We explore the use of Gaussian Processes for finite size extrapolations. The study and computation of material properties through ab initio electronic structure methods at large length scales usually becomes intractable because of excessive computational cost. As a result, reasonable results at large scales are routinely obtained by extrapolation of results at representative sizes. In order to obtain more reliable and efficient extrapolations, here, we use Gaussian processes [1]. We explore the use of Gaussian Approximation Potentials (GAP) [2] and Smooth Overlap of Atomic Positions (SOAP) [3] descriptors to extrapolate the total energy of smaller strongly correlated systems to estimate the energy of systems at larger sizes.

**Model**

The GAP approach expresses the energy as the sum of linear contributions of the local energy of the functional related to the $i$ atoms in the neighborhood some point in space, as in Eq. 1.

$$E_{total} = \sum_i \alpha_i \bar{K}(\{p_i\}, \{p_i^{\star}\})$$  \hspace{1cm} (1)

$\alpha_i$ are the descriptors of the local atomic environment on a structure and $\bar{K}(\{p_i\}, \{p_i^{\star}\})$ is the kernel which correlates the descriptors of different atomic environments. We use SOAP descriptors implemented in the Dcibe library [7] Eq. 2.

**SOAP DESCRIPTORS**

$$K(A, B) = \frac{1}{NM} \sum_{ij \in A \times B} k_i, j(A, B)$$ \hspace{1cm} (2)

We use the intuitive average kernel approach to compare atomic structures $A$ and $B$ with number of atoms $M$ and $N$ given a kernel $\hat{l}$. Eq. 3. This is equivalent to averaging the local descriptors over all of the atoms of each configuration Eq. 4.

**AVERAGE KERNEL**

$$\bar{K}(\{p_i\}, \{p_i^{\star}\}) = 1 + \sum_{ij}(\bar{K}(\{p_i\}, \{p_i^{\star}\}) - \hat{l})$$ \hspace{1cm} (3)

**MODEL KERNEL (RBF)**

We use the Radial Basis Function (RBF) kernel Eq. 5 and optimize its parameters to maximize the marginal log-likelihood [1]. We perform feature selection based on leverage scores for CUR decomposition to reduce the number of features similarly to the reference [9] but using hard-thresholding and without re-orthonormalization.

$$K(p, p') = \sigma^2 \exp(-\frac{1}{2} \frac{(\vec{r}(p, p') - \vec{r}(\bar{p}, \bar{p})))}{\bar{\lambda}^2})$$ \hspace{1cm} (4)

$$\hat{\sigma}, \bar{\lambda}$$ are the hyperparameters for Computation and Visualization, Brown University.

This system exhibits a metal-to-insulator transition when stretched homogeneously. The transition is continuous with respect to energy, but can be characterized by complex polarization or spin correlation functions [6]. Dimerization can be observed by looking at the electron density profile along the chains in Figure 1 (right). The features of the density profile can be used as order parameters that can characterize the transition [6].

**Results**

We train the GPR using chains of 10, 20, and 30 atoms, and mix these points in a random way to form the training set. We then predict the energy of chains of larger sizes. Figure 2 left shows data points generated with Hartree-Fock and UCCSD(T), compared against the GPR prediction. At the right, the difference between the UCCSD(T) and the GPR predicted energies per atom.

**References**

Rubenstein Group, Department of Chemistry, Brown University

**Acknowledgements**

The authors gratefully acknowledge financial support from the Research Corporation of America, U.S. National Science Foundation grant OIA-1921199, and U.S. Department of Energy, Office of Science, Basic Energy Sciences Award DE-FOA-0001912. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

Fininte Size Effects

Figure 3 displays the UCCSD(T) and GPR energies per atom against the inverse of the chain size, $N^{-1}$. Dashed lines represent a second order polynomial fit in $N^{-1}$. The convergence is slow or fast for small and large bond lengths, respectively.

**Figure 3:** Finite size effects on the energy per atom vs. $N^{-1}$. Cyan dots and maroon xs denote the UCCSD(T) and GPR predictions; the shadow depicts 95% confidence bounds, and the dashed lines depict the polynomial-regression of second order at each bond length.

Figure 4 displays the energies and their differences with respect to a reference for larger sizes. These energies approach the thermodynamic limit with chemical accuracy.

**Figure 4:** Left: Energy per atom for $N=150,200,\rightarrow \infty$, the UCCSD(T) reference energies were computed by us. The "REF $\rightarrow \infty$" was taken from [4], the GPR thermodynamic limit is computed using chains of 3000 atoms. Right: Difference of energy between the reference and predicted energies at the different sizes.

**Figure 5:** Finite size effects on the energy per atom vs. $N^{-1}$. Cyan dots and maroon xs denote the UCCSD(T) and GPR predictions; the shadow depicts 95% confidence bounds, and the dashed lines depict the polynomial-regression of second order at each bond length.

**Figure 6:** Left: Energy per atom for $N=150,200,\rightarrow \infty$, the UCCSD(T) reference energies were computed by us. The "REF $\rightarrow \infty$" was taken from [4], the GPR thermodynamic limit is computed using chains of 3000 atoms. Right: Difference of energy between the reference and predicted energies at the different sizes.