# Atom-in-molecule based quantum machine learning of defect formation energies

**Alastair James Arthur Price** [1 2] **O. Anatole von Lilienfeld** [1 3 2 4 5] **Stephen Dale** [6]

[1]*Department of Chemistry, University of Toronto, St. George campus, Toronto, ON, Canada* [2]*Vector Institute for Artificial Intelligence, Toronto, ON, Canada* [3]*Department of Materials Science and Engineering, University of Toronto, St. George campus, Toronto, ON, Canada* [4]*Laboratory for AI and automation, Acceleration Consortium, University of Toronto, Toronto, ON, Canada* [5]*Department of Physics, University of Toronto, St. George campus, Toronto, ON, Canada* [6]*National University of Singapore*. Correspondence to: Stephen Dale sdale@nus.edu.sg.

## 1. Introduction

Defects govern the electronic, optical, and mechanical properties of solid-state materials, from acting as single-photon emitters (SPEs) in hexagonal boron nitride (*h*-BN) to influencing thermal transport in thermoelectrics. While Density Functional Theory (DFT) provides the necessary accuracy for modeling defect energetics, its cubic scaling ($O(N^3)$) limits the exploration of the large supercells required to minimize periodic image interactions.

To address this, we apply an Atom-in-Molecule (Amons) based hierarchical approach.[1, 2] Unlike previous molecular implementations that rely on gas-phase clusters, we introduce a solid-state Amons ansatz. We define "amons" as a hierarchical set of small periodic supercells. By analyzing these computationally inexpensive fragments, we derive the relationship between finite-size scaling and energy. These trends are extrapolated to predict the formation energies of large, dilute-limit supercells (e.g., $17 \times 17$) that are computationally prohibitive to treat directly with DFT.

We validate this workflow on substitutional defects and vacancies in *h*-BN, graphene, and diamond, demonstrating that a simple linear scaling fit based on small fragments can recover bulk-limit properties with DFT-level accuracy.

## 2. Methodology

### *2.1 DFT Calculations*

Reference calculations were performed using the FHI-aims code[3] with numeric atom-centered basis functions (NAOs).[4] We employed the PBE functional[5] with exchange-hole diploe moment (XDM)[6, 7] dispersion corrections to capture layer-dependent properties. Geometries were relaxed using the BFGS algorithm until residual forces fell below $5.0 \times 10^{-3}$ eV/Å. To maintain consistency with the hierarchical Amons approach, a Γ-point only sampling was used.

### *2.2 Dataset and Extrapolation Scheme*

We constructed a hierarchical dataset of supercells ranging from $2 \times 2$ (high concentration) to $17 \times 17$ (dilute limit). The "Amons" are defined as the smaller periodic supercells, which preserve Periodic Boundary Conditions (PBC) and long-range lattice strain.

Rather than employing a high-dimensional many-body representation, we utilize a robust extrapolation based on system size. The formation energy $E_f$ is modeled as a function of the inverse supercell volume ($1/V$) or characteristic length ($1/L$):

$$E_f(V) = E_{\text{dilute}} + \frac{\alpha}{V} + \dots \tag{1}$$

By fitting this relationship to a series of small, inexpensive Amons, we extract the dilute limit energy $E_{\text{dilute}}$. This approach avoids the noise associated with complex descriptors on small datasets and directly targets the physical scaling law governing defect energetics.

## 3. Results and Discussion

### *3.1 Finite-Size Scaling and Convergence*

A key challenge in defect modeling is accurately capturing the energy convergence as the simulation cell grows toward the dilute limit. Our hierarchical model successfully captures this physical scaling law, allowing for extrapolation to the dilute limit (approx. $45,000$ Å$^3$).

To quantify the efficiency of this extrapolation, we analyzed convergence curves for the Nitrogen vacancy ($V_{\text{N}}$) in *h*-BN (Figure 1). The plot illustrates the error in the predicted dilute-limit energy as larger Amons are progressively included in the fit.

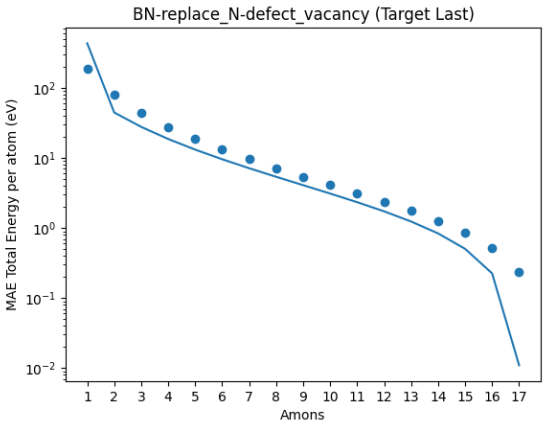


Fig. 1: **Convergence Analysis.** Prediction error (MAE) for the $V_{\text{N}}$ vacancy in a $17 \times 17$ cell as a function of the number of Amons included in the fit. The error decays rapidly, achieving sub-eV accuracy with minimal computational cost.

The Carbon substitution defect ($C_{\text{N}}$) converges even faster due to minimal lattice distortion. This

demonstrates that chemical accuracy in insulating 2D materials is achievable via simple scaling laws derived from a fraction of the computational cost of full supercell DFT.

### *3.2 Transferability to Graphene*

We extended the study to graphene to test performance on semimetals. While Boron substitution ($B_C$) converges smoothly, the Carbon vacancy ($V_C$) exhibits a more complex convergence curve due to delocalized $\pi$-electrons and long-range screening. However, by including slightly larger amons in the scaling fit, the model recovers the correct behavior, highlighting the robustness of the hierarchical approach.

## 4. Conclusion

We presented a solid-state Amons framework for predicting defect energetics via finite-size scaling. By utilizing a hierarchy of small periodic supercells, the approach inherently preserves boundary conditions and captures asymptotic energy convergence. We achieved high accuracy for *h*-BN and demonstrated transferability to graphene. Furthermore, the systematic consistency of these scaling laws suggests the generated data is highly amenable to future Quantum Machine Learning (QML) approaches, potentially serving as robust physical priors for $\Delta$-learning schemes.

Future work will leverage this efficiency to enable high-throughput screening using computationally demanding hybrid functionals (e.g., PBE0+XDM), which are currently prohibitive for large defect supercells.

## References

[1] Bing Huang and O Anatole von Lilienfeld. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nature chemistry*, 12(10):945–951, 2020.

[2] Bing Huang and O Anatole von Lilienfeld. The "dna" of chemistry: Scalable quantum machine learning with "amons". *arXiv preprint arXiv:1707.04146*, 2017.

[3] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Comp. Phys. Comm.*, 180:2175–2196, 2009.

[4] Ville Havu, Volker Blum, Paula Havu, and Matthias Scheffler. Efficient o(n) integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.*, 228:8367–8379, 2009.

[5] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865, 1996.

[6] E R Johnson. The exchange-hole dipole moment dispersion model. In A Otero-de-la-Roza and G A DiLabio, editors, *Non-covalent Interactions in Quantum Chemistry and Physics*, chapter 5, pages 169–194. Elsevier, 2017.

[7] A. J. A. Price, A. Otero de la Roza, and E. R. Johnson. Xdm-corrected hybrid dft with numerical atomic orbitals predicts molecular crystal lattice energies with unprecedented accuracy. *Chem. Sci.*, 14:1252–1262, 2023. Chapter **??** in this thesis.