Supplementary material: Solving the electronic structure problem with machine learning

Anand Chandrasekaran,[1] Deepak Kamal,[1] Rohit Batra,[1] Chiho Kim,[1] Lihua Chen,[1] Rampi Ramprasad[1]*

[1]School of Materials Science and Engineering, Georgia Institute of Technology,
771 Ferst Dr. NW, Atlanta, GA 30332, USA

*To whom correspondence should be addressed; E-mail: rampi.ramprasad@mse.gatech.edu

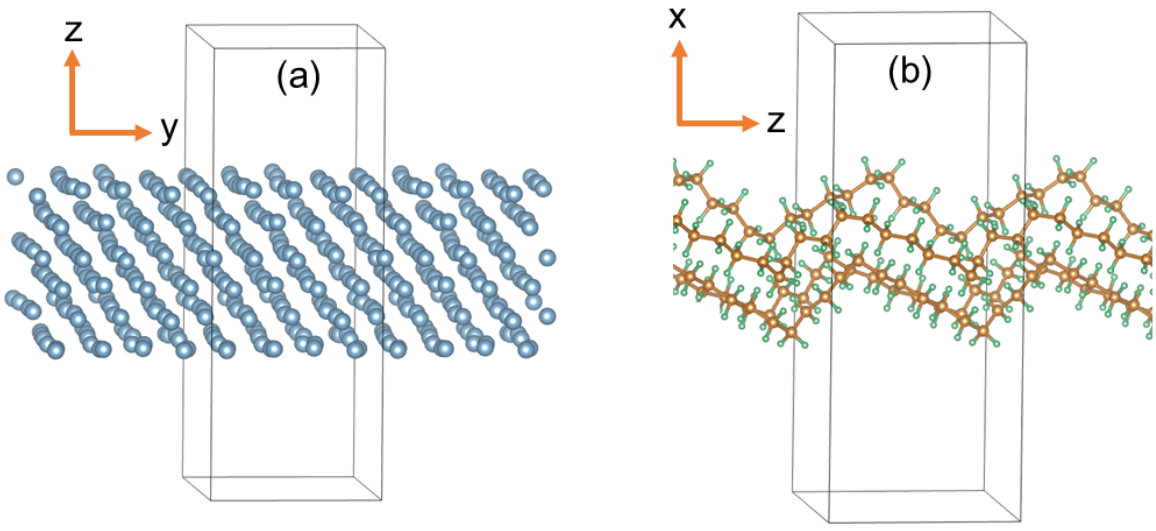

Figure S1: Typical structure of one of the snapshots of (a)Al (001) slab and (b) PE slab considered in this study.

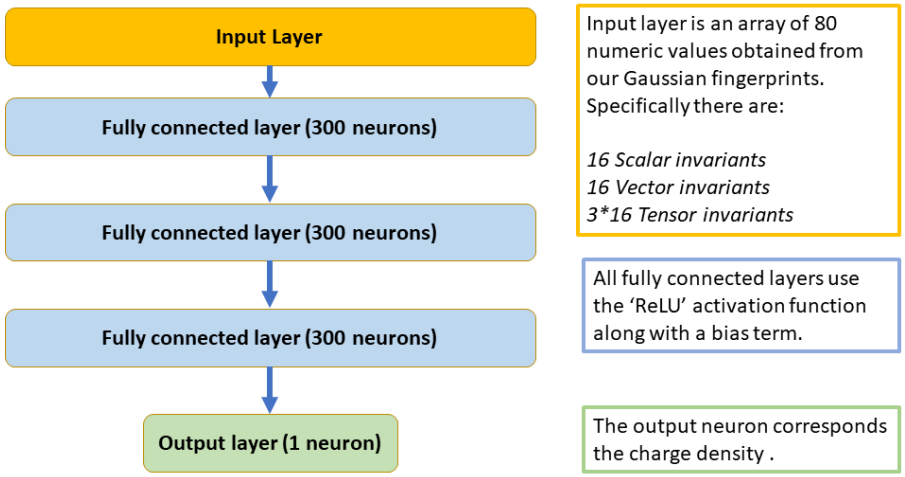

Figure S2: Neural network architecture for the training and prediction of charge density.

Table S1: Train, test and validation error metrics for polyethylene snapshots.

	Average RMSE	**Max 1000 RMSE**	**R^2**	**R^2**
	e/Å^3	**e/Å^3**	**Charge density**	**DOS**
Train snapshot 1	0.00037	0.00615	0.9999976	0.99913
Train snapshot 2	0.00038	0.00634	0.9999976	0.99833
Train snapshot 3	0.00040	0.00739	0.9999973	0.99833
Train snapshot 4	0.00038	0.00627	0.9999975	0.99778
Train snapshot 5	0.00038	0.00649	0.9999976	0.99821
Train snapshot 6	0.00038	0.00646	0.9999976	0.99768
Train snapshot 7	0.00038	0.00659	0.9999975	0.99859
Train snapshot 8	0.00037	0.00638	0.9999976	0.99787
Train snapshots average	0.00038	0.00652	0.9999975	0.99824
Validation snapshot	0.00040	0.00666	0.9999973	0.99817
Test snapshot	0.00037	0.00627	0.9999976	0.99768

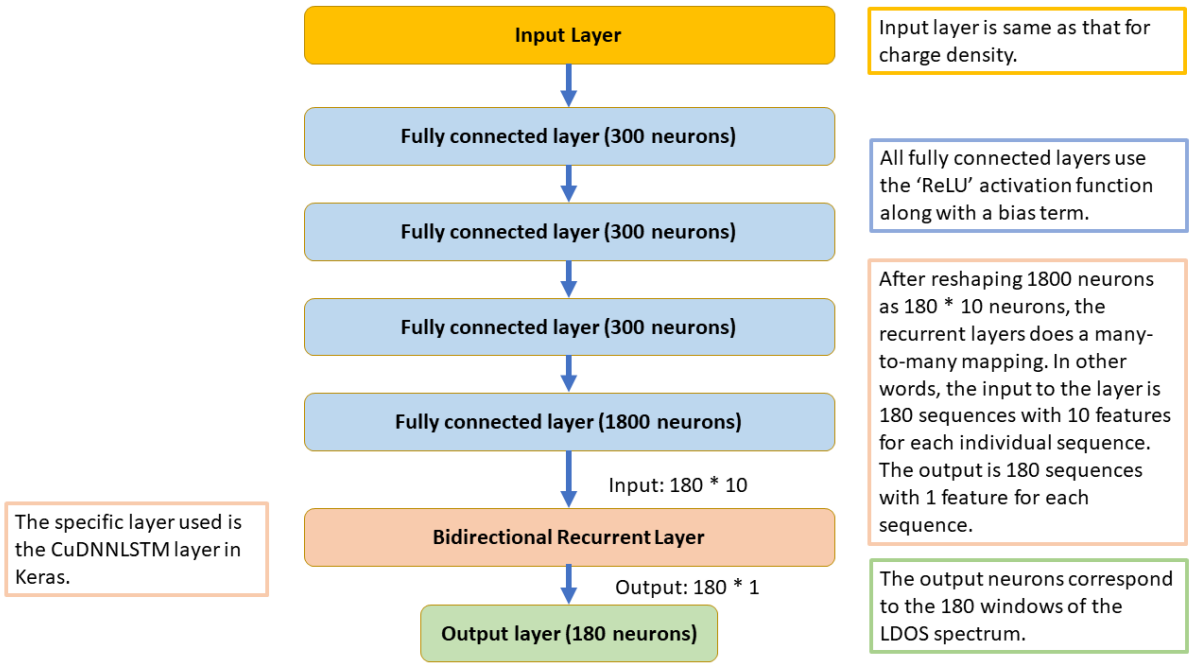

Figure S3: Neural network architecture for the training and prediction of the 180 windows of the LDOS spectrum of aluminum.

Table S2: Train, test and validation error metrics for aluminum snapshots.

	Average RMSE	**Max 1000 RMSE**	**R^2**	**R^2**
	eÅ^3	**e/Å^3**	**Charge density**	**DOS**
Train snapshot 1	0.00063	0.00438	0.999947	0.9993
Train snapshot 2	0.00061	0.00419	0.999951	0.9995
Train snapshot 3	0.00059	0.00412	0.999953	0.9994
Train snapshot 4	0.00058	0.00421	0.999956	0.9993
Train snapshot 5	0.00058	0.00422	0.999955	0.9994
Train snapshot 6	0.00059	0.00426	0.999954	0.9993
Train snapshot 7	0.00061	0.0044	0.999950	0.9993
Train snapshot 8	0.00065	0.00463	0.999944	0.9993
Train snapshots average	0.00061	0.0043	0.999951	0.9994
Validation snapshot	0.00059	0.00414	0.999954	0.9992
Test snapshot	0.00058	0.00419	0.999955	0.9992

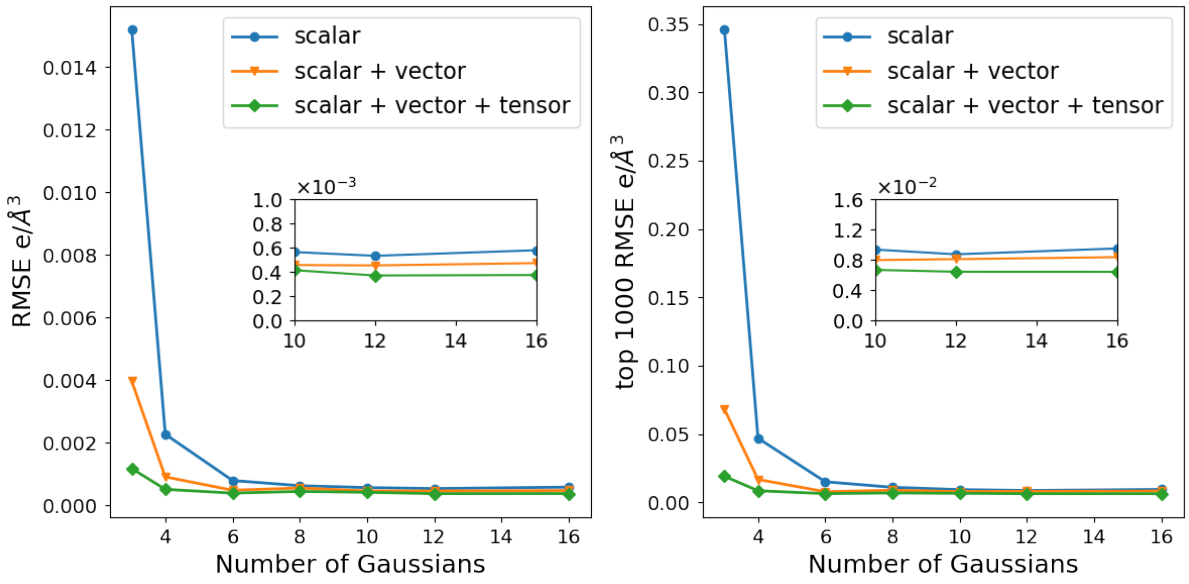

Figure S4: Average total RMSE and the average RMSE of 1000 points with the highest error for the case of the PE test snapshot. A systematic improvement is observed when increasing the number of Gaussians and also when including the vector and tensor invariants.

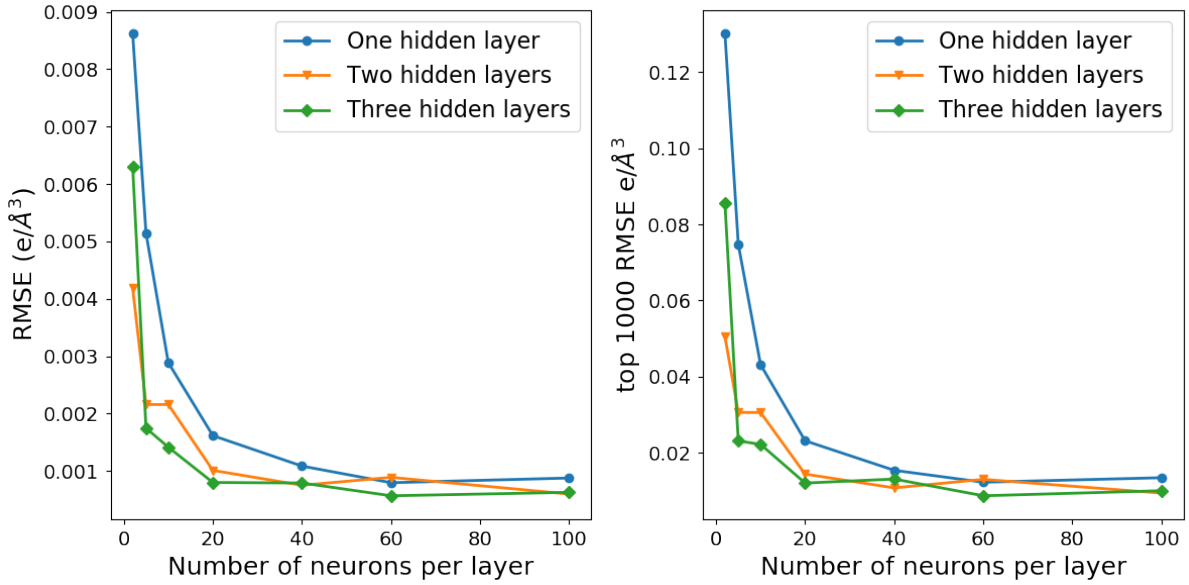

Figure S5: Average total RMSE and the average RMSE of 1000 points with the highest error for the case of the PE test snapshot. A systematic improvement is observed when increasing the number of hidden layers and number of neurons per hidden layer.

1 ML Charge density prediction - Basis-set approach vs grid-based approach

A different approach to by-pass the Kohn-Sham equations has recently been reported by Brockherde et al. (*1*). Their approach involves the representation of the charge density in terms of a basis set followed by the utilization of machine learning (ML) to predict the basis set coefficients. As highlighted in Figure. 4 of their work (*1*), the basis set representation introduces far more error than the actual ML mapping itself. By circumventing the basis set representation and directly predicting the charge density at grid points we are able to improve the fidelity of the machine learning prediction. The training set was generated from 30 molecular dynamics snapshots of a benzene molecule obtained from http://quantum-machine.org/datasets/. Specifically the 30 snapshots were taken randomly from the file "benzene_300K-400K.tar.gz". DFT calculations were performed on these structures to obtain the charge density. Once the charge-density model was constructed from these snapshots we tested our model on the 200 snapshots provided in the file "benzene_300K-test.tar.gz". By comparing Figure S6(b) with Figure 4b in the work by Brockherde et al. (*1*), we see an improvement in the accuracy of the charge-density prediction using our grid-based ML scheme. Moreover, a further benefit of the grid-based approach is that such levels of accuracy were achieved using just 30 training snapshots rather than 2000 training snapshots (*1*).

However, since the number of grid-points are usually larger than the number of basis-set coefficients, this improved accuracy is likely achieved at the cost of increased computational time. Our vectorized implementation on GPUs offsets this computational cost to a large extent but the relative merits of the basis-set approach vs the grid-based approach will be examined more closely in a later work.

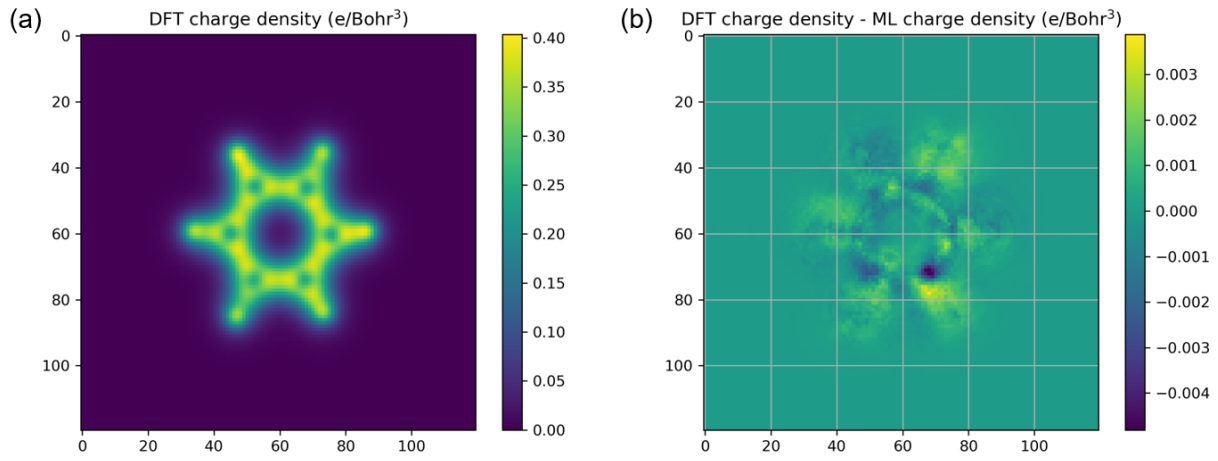

Figure S6: (a)DFT charge density of a molecular dynamics snapshot of a benzene molecule obtained from the website http://quantum-machine.org/datasets/. (b) Charge density difference between ML prediction and DFT for the same benzene structure.

2 Calculating energies from ML-predicted charge density

Although both the charge density and DOS ML models show very high accuracy, even small changes in the charge density may result in large fluctuations of the total energy. To investigate the effect of small fluctuations of the charge density on the total energy, we first randomly perturbed the DFT charge density at every grid-point for the 200 MD test structures of benzene (mentioned in the previous section). In other words, the magnitude of the charge density at every grid-point was perturbed randomly by small percentage of its initial value (the magnitude of perturbation was between 0% and 1% of the actual value). This perturbed charge density was subject to a non-self-consistent electronic minimization step (keeping the charge density constant) in order to obtain the total energy. The thus obtained total energy was compared to the actual DFT charge density obtained fully self-consistently as shown in Figure S7. The root-mean-square-error of the total energy of the 200 test snapshots using this slightly perturbed charge density was 45 meV (or 1.05 kcal/mol). On the other hand, when we use the *ML-predicted charge density* as the starting point for the non-self-consistent calculation, we obtained

an RMSE of just 2.8 meV (0.065 kcal/mol) and a maximum absolute error of 20.8 meV (0.48 kcal/mol).

Although these results prove that our charge density predictions are indeed accurate enough to provide highly accurate estimates of the total energy, a more thorough investigation of the utilization of both the charge density and DOS to *directly* obtain the total energy (via Equation 4 in the manuscript) will be pursued in a future work.

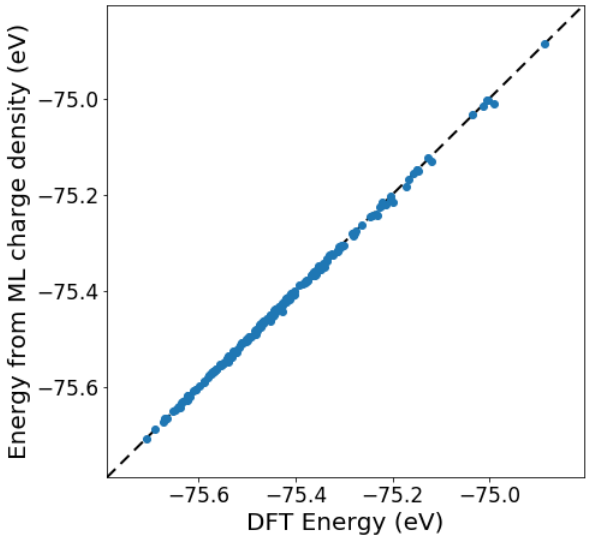

Figure S7: Parity plot of total energies for benzene molecule MD snapshots (from the test-set) obtained (non-self-consistently) using the ML charge density vs the total energies obtained (self-consistently) from DFT.

References

1. Brockherde, F. *et al.* Bypassing the Kohn-Sham equations with machine learning. *Nature Commun.* **8**, 872 (2017).

Table S3: Scaling tests comparing DFT (VASP) vs the machine learning algorithm for the prediction the charge density and DOS for various supercells of Al (100) slabs.

Number of atoms		144	576	1296	1728	2304
Number of grid-points (million)		5.9	23.3	52.7	70.3	93.6
DFT	Computational time (s)	860	13010	50442	106458	Memory error
	Scaling constant	-	2.0	1.9	1.9	-
ML	Fingerprinting (s)	23	78	168	222	296
	Charge density prediction (s)	5	20	46	60	90
	DOS prediction (s)	11	41	96	132	193
	Total time (s)	49	165	366	490	685
	Scaling constant	-	0.9	0.9	0.9	1.0
DFT time / ML time		18	79	138	217	-