Supporting Information for Unsupervised Representation Learning of Kohn-Sham States and Consequences for Downstream Predictions of Many-Body Effects

Bowen Hou[1], Jinyuan Wu[1], and Diana Y. Qiu*[1]

[1]Department of Mechanical Engineering and Materials Science, Yale University, New Haven, CT, 06511, USA

October 12, 2024

In this document, we supplement the statements in the main text by discussing details of the VAE model, training dynamics, application to ground-state properties, GW calculations, and the pseudobands approximation.

Supplementary Note 1—VAE model

Here, we will mathematically demonstrate that the VAE latent space from the special designed encoder can effectively represent the KS state of a specific material with a unique unit cell, and what the latent space exactly learns. Suppose we want to learn the representation of a 2D KS state:

$$|\phi_n(x,y,c)|_{I_{(x,y\in(-\infty,+\infty))}} \tag{1}$$

with periodicities T_x and T_y along two in-plane lattice constant vectors. Due to the 2D nature, we treat the z degree of freedom of the KS states as an input channel c. n represents the quantum number and crystal momentum of a specific material. However, instead of working in infinite space, we typically only have the data from one complete periodicity within a unit cell:

$$f_n(x,y,c;t_x,t_y) = |\phi_n(x+t_x,y+t_y,c)|_{I_{(x\in(0,T_x),y\in(0,T_y))}} \tag{2}$$

where arbitrary translation t_x and t_y is the unphysical degrees of freedom from the choice of the origin of the unit cell. This results in a challenge for regular CNNs-NN architecture because the translational covariance cannot be preserved by passing CNN output to a dense layer, even if the CNN itself has translational covariance. In our context, to achieve prediction with physical invariance, the first crucial point is to introduce circular padding to enable the periodic boundary condition and ensure the periodicity of the CNN output feature map is the same as the input layer. This way, the padding width equals the kernel size so that the convolutional scanning will always go through the periodicity of the input. Then, we can obtain the feature map by inputting f_n to CNNs followed by a non-linear layer:

$$\tilde{O}_n(X,Y,C';t_x,t_y) = \mathrm{ReLU}\left(\sum_c \sum_x^{T_x} \sum_y^{T_y} f_n(X+x,Y+y,c;t_x,t_y)\cdot K(x,y,C')\right) \tag{3}$$

where C' is the output channel number, and the nonlinear ReLU function is:

$$\mathrm{ReLU} = \frac{x+|x|}{2} \tag{4}$$

Importantly, the output feature map $\tilde{O}_n(X,Y,C';t_x,t_y)$ not only provides additional nonlinear degrees for fitting/representation, but it also preserves the periodicity of T_x and T_y introduced by circular padding. Then, to achieve translational invariance, we utilize the basic fact that the integral of a periodic function over a complete periodicity is always invariant to any arbitrary

*Corresponding author: diana.qiu@yale.edu

shift t_x and t_y. Therefore, instead of directly passing $\tilde{O}_n(X, Y, C'; t_x, t_y)$ to the next layer for learning, we sum the feature map over X, Y, and we use $\tilde{a}_n(C')$ to denote the summation:

$$\begin{aligned}\tilde{a}_n(C') &= \sum_{X}^{T_x}\sum_{Y}^{T_y}\tilde{O}_n(X, Y, C'; t_x, t_y) \\ &= \sum_{X}^{T_x}\sum_{Y}^{T_y}\tilde{O}_n(X, Y, C')\end{aligned} \tag{5}$$

where the unphysical degree of freedom of t_x and t_y are summed out with X and Y, which achieve physical invariance to arbitrary translations . As a result, the consequent output vector $\tilde{\mathbf{a}}_n \in \mathbb{R}^{C_{\text{out}}}$ through the global channel aggregation can be used as representations for f_n if C_{out} is large enough. The resulting loss of local spatial information of X and Y can be totally compensated by the depth of the channels, which provides sufficient nonlinearity for accurate fitting.

Additionally, we find that the above representations can also be used for large systems beyond single unit cell with minor modifications. Suppose we consider the KS state of an $M \times N$ super cell as:

$$\begin{aligned}F_n(x, y, c; t_x, t_y) &= |\phi_n(x + t_x, y + t_y, c)|_{I_{(x\in(0,N\times T_x),y\in(0,M\times T_y))}} \\ &= \sum_{i}^{N}\sum_{j}^{M} f_n(x + i \times T_x, y + j \times T_y, c; t_x, t_y)\end{aligned} \tag{6}$$

Passing this to our CNN model mentioned above, we get:

$$\begin{aligned}\tilde{A}_{M,N,n}(C') &= \sum_{i}^{N}\sum_{j}^{M}\sum_{X}^{T_x}\sum_{Y}^{T_y}\tilde{O}_n(X + i \times T_x, Y + j \times T_y, C') \\ &= (M \times N)\sum_{X}^{T_x}\sum_{Y}^{T_y}\tilde{O}_n(X, Y, C') \\ &= (M \times N)\tilde{a}_n(C')\end{aligned} \tag{7}$$

where the second line comes from the integral invariance of a periodic function. Intriguingly, the representation of an $(M \times N)$ super cell $\tilde{A}_{M,N,n}(C')$ is always equal to $(M \times N)$ times that of a unit cell $\tilde{A}_{M=1,N=1,n}(C')$. Since our model doesn't require any cropping/resizing for the preprocessing, the original cell size information is preserved properly. Then, we can average each nonlinear feature map (essentially the nonlinearly curved density of filtered reciprocal space components) from the original electron state as follows:

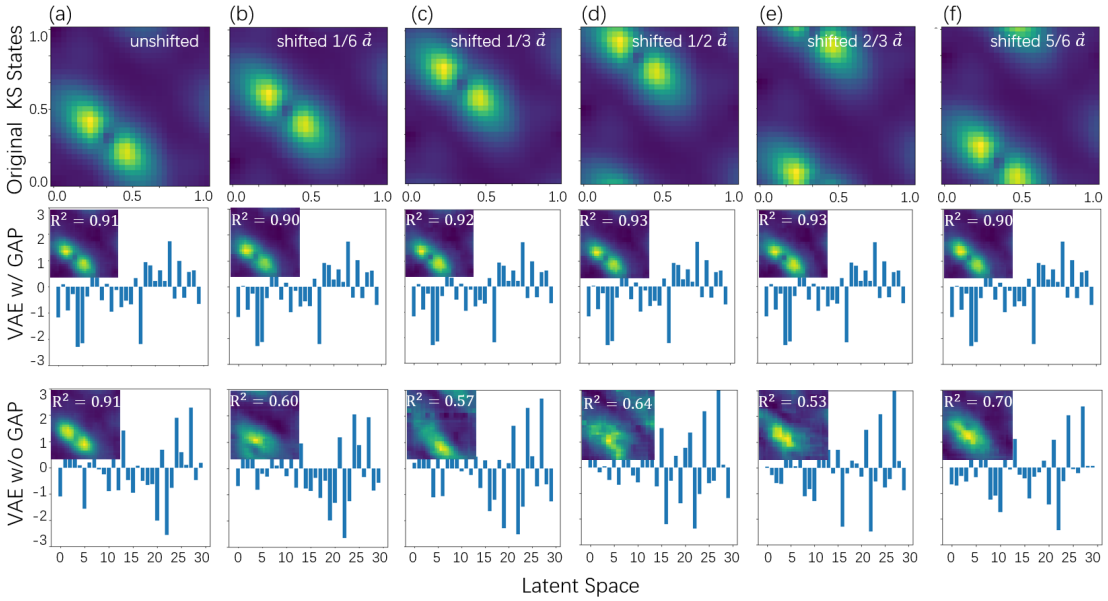
$$\begin{aligned}\mathbf{z}_{M,N,n}(C') &= \frac{1}{NT_x \times MT_y}\tilde{A}_{M,N,n}(C') = \frac{1}{T_x \times T_y}\tilde{a}_n(C') \\ &= \frac{1}{T_x \times T_y}\tilde{A}_{1,1,n}(C') \\ &= \mathbf{z}_{1,1,n}(C')\end{aligned} \tag{8}$$

As a result, the latent space $\mathbf{z}_n = f(\mathbf{z}_{1,1,n}(C'))$ (f is following dense layer) remains the same for any supercell size, and it can represent a complete KS state density $|\phi_n(x, y, c)|_{I_{(x,y\in(-\infty,+\infty))}}$ in the whole space. More importantly, our model is never limited to the data within a unit cell spanned by two vectors with a specific angle. Instead, given the same material, the data from any complete periodicity of a specific KS state will generate the same $\mathbf{z}_{1,1,n}(C')$. Therefore, the symmetry of the unit cell is implicitly enforced even though input data does not include explicit symmetry labels, and the grid-based data can be properly used as the input layer if a continuous periodicity is ensured. Here, $\mathbf{z}_{1,1,n}(C')$ can be interpreted as a vector of "average charge density" from different nonlinear global feature maps of a state.

In summary, our VAE model has several advantageous properties to generate effective representations: i) $\mathbf{z}_n$ is always smooth due to the smooth nature of neural networks, ii) $\mathbf{z}_n$ can handle translational invariance and PBC, iii) can handle any cell size and symmetry, iv) even if

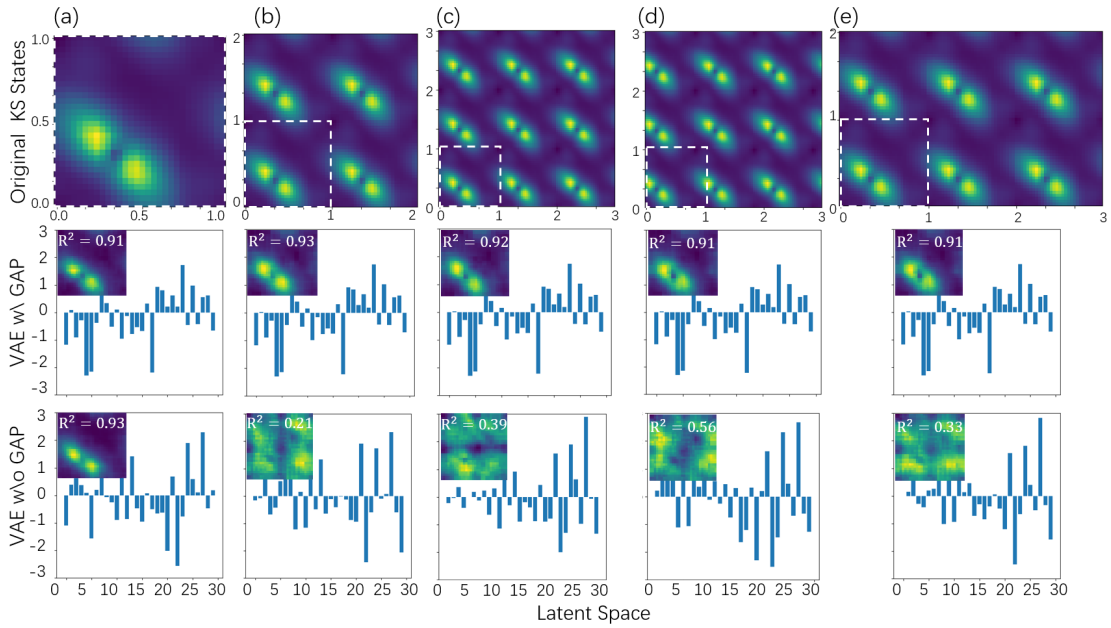
it is only trained with unit cell data, $\mathbf{z}_n$ can be extended to large systems beyond the unit cell, which provides future opportunities to study systems like defects and moiré bilayers.

Next, we demonstrate additional generalizability of our model. In the first row of Supplementary Supplementary Fig. 1(a-f), we select one random KS state from the test set and gradually slide the window of the unit cell along one of its lattice constant vectors. The second row of Supplementary Supplementary Fig. 1(a-f) displays the VAE latent space and reconstructed KS wavefunction, obtained by inputting the corresponding shifted wavefunction into our model. The generated wavefunction remains entirely invariant to any sliding of the unit cell. The third row demonstrates the latent space and generated wavefunction using a model without including GAP, which is sensitive to the selection of the unit cell. We see that the VAE with modified CNN layer significantly enhances our model's ability to deal with systems of different unit cell size, cutoff energies, PBC and translational invariance.



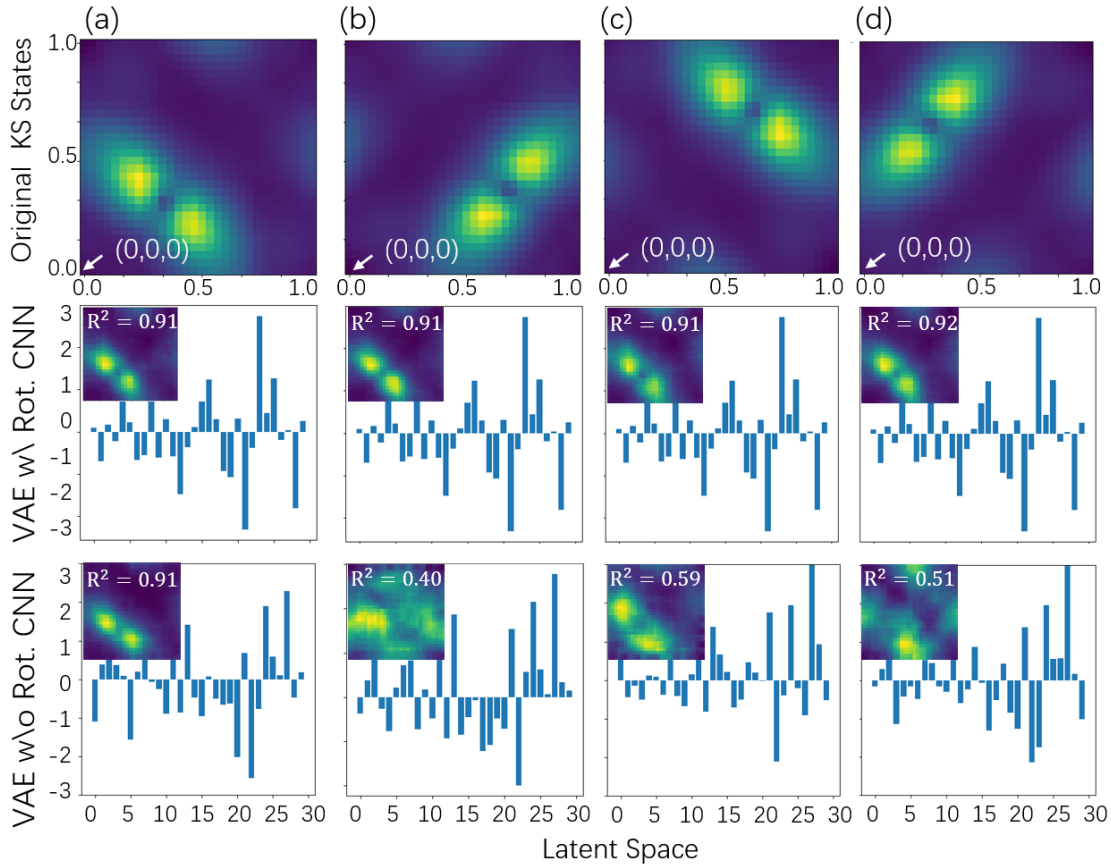
Supplementary Figure 1: The first row of (a-f) shows a randomly picked KS state shifted along the y-direction with 0, $\frac{1}{6}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{2}{3}$, and $\frac{5}{6}$ **a** in real space. The second row of (a-f) illustrates the latent space of the related shifted KS states respectively. The insets are the corresponding reconstructed wavefunctions, generated by decoding the latent space. The latent space (reconstructed wavefunction) is invariant to the unphysical degree of freedom of choosing the original point of the unit cell, which preserves the translational invariance for the downstream physical prediction. The third row of (a-f) is the latent space and reconstructed KS states by a simple VAE without including GAP layer, which are presented for comparison.

As Supplementary Fig. 2 illustrates, we also show that our model is independent of the size of the input in real space and can be extended to large super cells:



Supplementary Figure 2: The first row of (a-e) illustrate the 1×1, 2×2, 3×3, 3×3 (shifted), and 2×3 super cell of a randomly picked KS state. The second row of (a-e) show the respective latent space of related KS states from different super cells. The insets are the corresponding reconstructed wavefunction, generated by decoding the latent space of rotated KS states. The reconstructed wavefunction (latent space) are invariant to the size of the super cell, which can be treated as a fundamental representation for KS state of the whole space. The third row of (a-e) are reconstructed KS states by a simple VAE without including GAP, which are presented for comparison.

Lastly, to investigate how our model handles the freedom of selecting the coordinate basis, we choose four combinations of lattice vectors $(\mathbf{a}_1, \mathbf{a}_2)$, $(\mathbf{a}_1, -\mathbf{a}_2)$, $(\text{-}\mathbf{a}_2, \mathbf{a}_1)$, and $(\mathbf{a}_2, -\mathbf{a}_1)$for the input crystal unit cell. These combinations correspond to 90, 180, and 270-degree rotations of the wavefunction in fractional crystal coordinates, as shown in the first row of Supplementary Fig. 3(a-d). The second row of Fig. Supplementary Fig. 3(a-d) illustrates the latent space and reconstructed wavefunction corresponding to the different choices of unit cell basis through our VAE and a VAE without rotational CNN (third row). As expected, integrating rotatinal CNN allow our VAE to recover the original pattern, while the naive VAE based on traditional CNN is sensitive to the unphysical choice of coordinate basis.



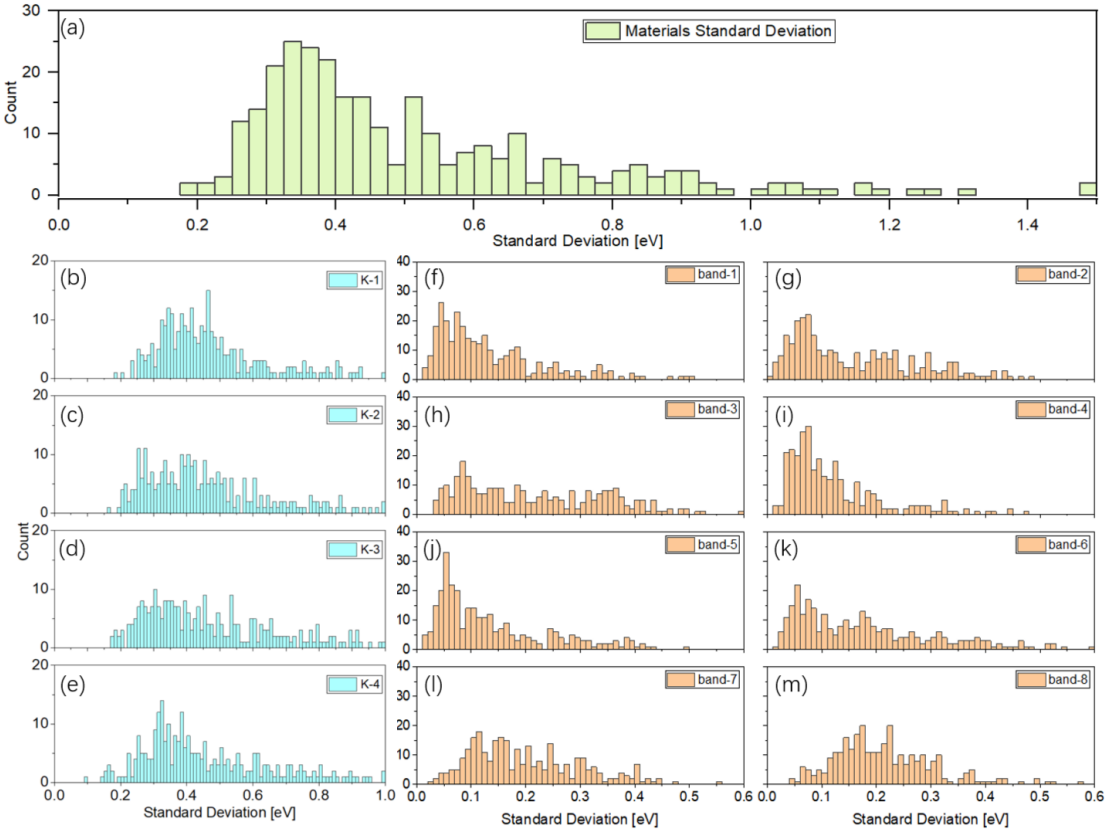
Supplementary Figure 3: The first row of (a-d) illustrates a randomly picked KS state represented in a unit cell with bases of lattice vectors $(\mathbf{a}_1, \mathbf{a}_2)$, $(\mathbf{a}_1, -\mathbf{a}_2)$, $(-\mathbf{a}_2, \mathbf{a}_1)$, and $(\mathbf{a}_2, -\mathbf{a}_1)$. The second row of (a-d) illustrates the latent space of respective rotated KS states. The insets are the corresponding reconstructed wavefunctions, generated by decoding the latent space of rotated KS states. The latent space (reconstructed wavefunction) is invariant to the choice of the lattice vectors. The third row of (a-d) shows reconstructed KS states by a naive VAE without including rotational CNN, which are presented for comparison.

Supplementary Note 2–Training Dynamics

Firstly, we want to emphasize the considerable challenge of learning the GW correction for the full bandstructure, even when the DFT bandstructure is known, which has been a longstanding problem in the field. Previous GW machine learning models, have relied on manually selected intermediate physical quantities based on human intuition. These early approaches were only capable of predicting the band gap [1, 2] or after carefully tailoring feature, were able to predict a k-point-dependent GW correction that lacked the smoothness of a physical bandstructure [3]. This suggests that learning the GW correction is inherently a hard problem highly sensitive to feature selection. In our work, we overcome this challenge by replacing feature selection with autonomous representation learning, allowing us to obtain smooth GW bandstructures for the first time. More fundamentally, unlike previous work that carefully tailored feature selection for the GW problem specifically, our representation learning is not designed to target GW, and it could be easily generalized in the future to any downstream method relying on DFT wavefunctions.

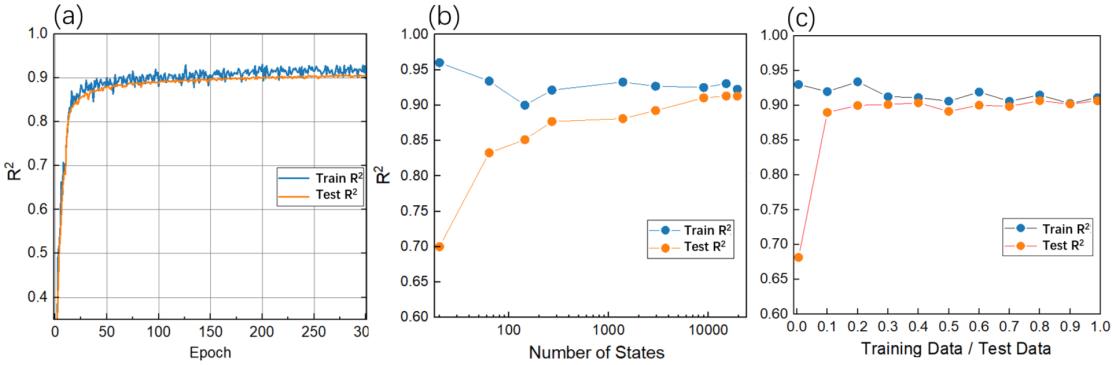
To quantitatively analyze the GW dataset and estimate the training difficulties, we investigate how the GW corrections vary with different k points and bands in each material in our test set. Firstly, we calculate the standard deviation of GW corrections for all bands and k-points within each material. Supplementary Fig. 4(a) presents the histogram of these standard deviations across all 302 materials. The standard deviation of GW corrections in each material is significantly larger than 0.1 eV, which is the MAE of our model. In addition, Supplementary

Fig. 4(b-e) shows the standard deviation of GW corrections across all bands for a given k point, plotted as a histogram for all 302 materials in the dataset. The standard deviations for each material significantly exceed 0.1 eV as well. Similarly, Supplementary Fig. 4(f-m) depict the histogram of standard deviations of GW corrections for all k points given the same band across all 302 materials, demonstrating that the standard deviation for the same band in many materials is also much larger than 0.1 eV, indicating that our model captures this variability. Therefore, neither our model nor the GW correction in general is reduced to a single value regression (i.e. a scissor shift).



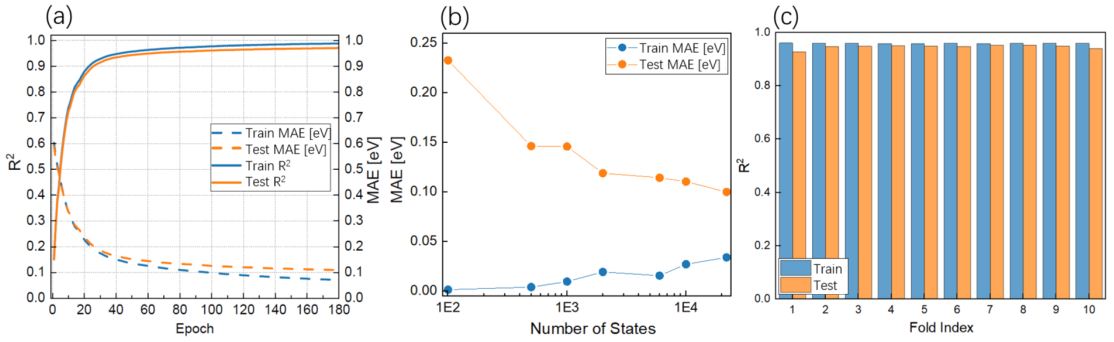
Supplementary Figure 4: (a) the histogram of standard deviations of GW corrections from all bands and k-points across all 302 materials. (b-f) the histogram of standard deviation of GW corrections from all bands given the same k-point across all 302 materials. (f-m) the histogram of standard deviation of GW corrections from all K-points given the same band across 302 materials. Source data are provided as a Source Data file.

Here, we provide more details regarding the training dynamics of both VAE and GW regression model. We first examined how the training process of the VAE progresses with the number of epochs. The total wavefunction data across 302 materials are randomly split, with 20% allocated to the test/validation set and the remaining 80% to the training set. The Supplementary Fig. 5(a) below shows the R^2 performance of the model on the training and validation sets, achieving high $\mathbf{R^2}$ values of 0.93 and 0.91 within 300 epochs, respectively. Additionally, to assess how the dataset size affects the VAE, we plot the training process with an increasing number of data points (KS states), as shown in Supplementary Fig. 5(b). Overfitting only appears when the number of data points is smaller than 1,000 states, while both the training and test sets achieve high R^2 values of 0.93 and 0.91 when the number of data points exceeds 10,000. To further verify this finding, we include 20,000 states in the data set, and change the ratio between training points and testing points. As Supplementary Fig. 5(c) shows, the overfitting is negligible when the training set is more than 30% (6,000 states), which is approximately consistent with previous result. In summary, this investigation into the training dynamics strengthens our statement that the VAE can provide general and efficient representations across different materials.



Supplementary Figure 5: (a) Training dynamics of VAE for KS states with respect to different number of epochs. (b) VAE R^2 performance with respect to different size of dataset. (c) VAE R^2 performance with different ratio of training/test data. The total dataset includes 20,000 KS states. Source data are provided as a Source Data file.

For the downstream supervised prediction, we also conduct more benchmarks on our model. Supplementary Fig. 6(a) shows that a simple downstream regression model achieves fast training convergence (180 epochs) with a low GW correction MAE of 0.06 eV on the training set and 0.11 eV on the test set. This is under our expectation due to the effective dimension reduction of the original wavefunction through the VAE. Supplementary Fig. 6(b) illustrates how the number of data points affects the model's performance, with overfitting becoming evident when the number of states is less than 1K. This issue is resolved by increasing the number of states to more than 10K. Additionally, we benchmark our model's performance using a stricter and more comprehensive cross-validation technique[4], as shown in Supplementary Fig. 6(c). Here, we randomly divide our dataset into 10 folds, then use the data from a specific fold to score the model trained on the remaining data. We iterate this process for all folds, and the average $\mathbf{R^2}$ score of our model is 0.96 on the training set and 0.94 on the test set.

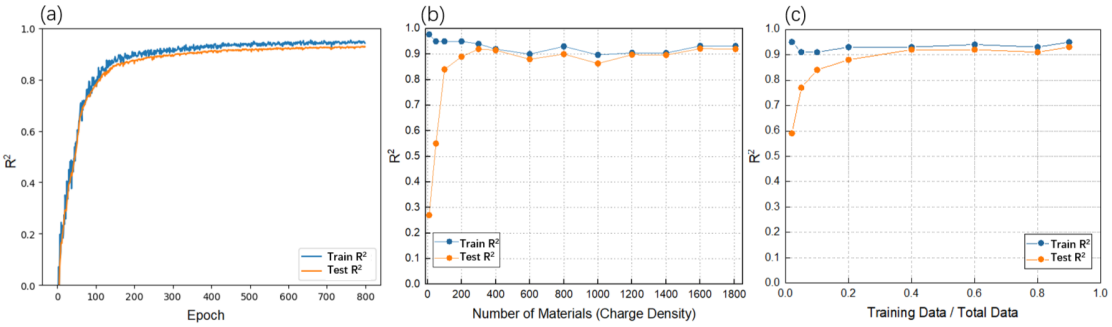


Supplementary Figure 6: (a) Training dynamics of downstream NN for GW corrections with respect to different number of epochs. The orange (blue) dashed lines represent test (training) MAE. The orange (blue) lines represent test (training) R^2. (b) Downstream NN R^2 of performance for GW prediction with respect to different size of dataset. (c) 10-folds cross-validation of downstream NN for GW prediction. Source data are provided as a Source Data file.

We also note that a few large outliers existing in our regression predictions. To identify these outliers, we filtered out materials and states with GW corrections larger than 0.25 eV. Although these account for only 4% of the test set, they contribute significantly to the final error. Despite the model's robust performance on both the training and testing sets, we observed that the outlier materials vary depending on some factors such as the choice of training set and model parameters. Therefore, the identification of a group of specific materials are not quite meaningful given this situation. A possible explanation for this variability is that some GW corrections may not be fully converged, resulting in the "noise" that which is learnt by model and leads to outliers.

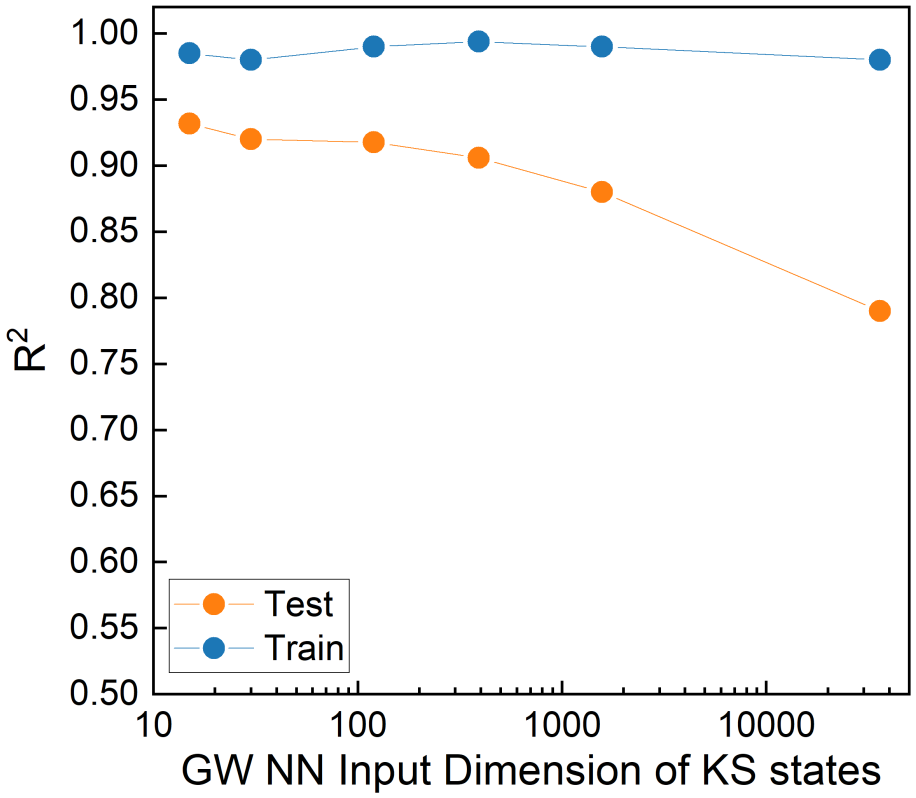
In our workflow, the representations of charge density, learned via a separate VAE (referred to as VAE-charge), are used as the inputs for the GW neural network. The charge density

representations can also be utilized for various downstream predictions of ground state DFT properties, which will be discussed in the following section. To train the VAE-charge, the total charge density data across 302 materials was randomly split, with 20% allocated to the test/validation set and the remaining 80% to the training set. Supplementary Fig. 7(a) shows the R^2 performance of the model on the training and validation sets, achieving a high R^2 value of 0.95 and 0.93 respectively. To assess the requirement of dataset and training dynamics of the VAE-charge model, we plot the training process with an increasing number of data points (charge densities of materials), as shown in Supplementary Fig. 7(b). Both the R^2 performance of the model on the training and testing sets exceeds 0.9 when the dataset includes 300 or more materials. Only when the training data points is smaller than 100, the model starts exhibiting overfitting. Supplementary Fig. 7(c) shows the R^2 performance of the model with respect to training/dataset ratio. These training dynamics demonstrate that our VAE model has strong extrapolation ability across different materials, supporting our claim of a general representation mentioned in the manuscript.



Supplementary Figure 7: (a) Training dynamics of VAE-charge with respect to different number of epochs. (b) VAE-charge R^2 performance with respect to different size of dataset. (c) VAE R^2 performance with different ratio of training/test data. Total dataset includes 2600 2D materials (charge density). Source data are provided as a Source Data file.

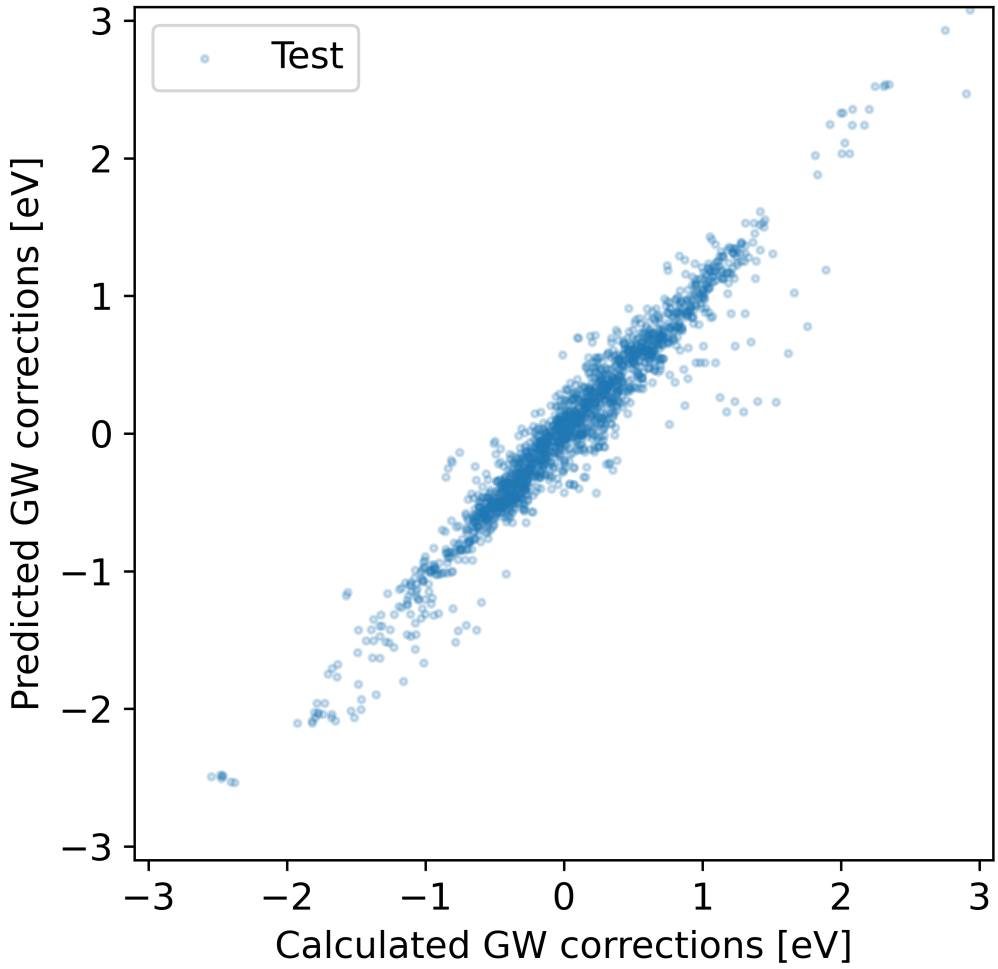
Next, we investigate how the data compression degree impacts the regression effect. If we directly apply neural networks (NN) to predict GW corrections, the network size becomes extremely large due to the extensive input layer. For instance, including Kohn-Sham (KS) states, charge density, and 10 super states in the input layer results in 39,602,801 trainable parameters in downstream NNs. The VAE reduces the total trainable parameters to 13,801. The large dataset before compression lead to several consequences: i) The training process for many-body predictions results in possible divergence. In our case, the direct prediction of GW self-energy by NN diverges within the first 3 epochs. ii) Overfitting is evident when data compression is not applied. To quantitatively analyze how data compression affects prediction accuracy, we plot the relationship between the R^2 of the GW regression and the compressed input size of the downstream neural network in Supplementary Fig. 8 . Our results indicate that overfitting decreases with higher levels of data compression. iii) High training memory requirement. For example, the DFT wavefunction across 302 materials is, in our case, around 1 TB, which will tremendously limit the batch size. iv) Training speed and convergence are significantly impacted by the small batch sizes, leading to extremely low learning efficiency. We observed that using data compressed to 1/1200th of its original size increases training speed by 500+ times compared to using the original data.



Supplementary Figure 8: R^2 of GW prediction with respect to different input dimensions of KS states. The orange (blue) dots represent model performance on test (training) set. Source data are provided as a Source Data file.

Finally, we look into the impact of data splitting methods on the training process for both the unsupervised VAE learning of DFT states and the supervised learning of GW correction. Rather than splitting the dataset based on states, we choose to segregate it by materials, reserving 30 distinct materials for testing purposes. For VAE learning, the performance of reconstructed wavefunctions in the test set is very robust to different batches of 30 randomly chosen materials in the test set. The r^2 of reconstructed wavefunction can achieve a high value of 0.92, which is almost the same level as the training set performance. However, for supervised learning, this approach revealed that GW prediction accuracy exhibits some sensitivity to the selection of different batches of materials, leading to a decline in the overall R^2 to a range of 0.80-0.90 (MAE≈0.17 eV) when material sets are chosen randomly.

We use four valence and four conduction bands across a fixed $6 \times 6 \times 1$ k-point grid to train the NN model. These states are chosen for convenience in generating training data and are irrelevant to the input/output size of the NN model. This choice will not introduce limitations related to training on a specific grid size. Once the f_{NN} model is well trained, it can take the latent space representation of any arbitrary KS state (n, k) and predict the GW correction for that specific KS state. As shown in Fig. S3(c), training the model at a specific k-point enables it to predict other nearby k-points accurately, demonstrating the model's extrapolation ability across arbitrary k-points. Furthermore, to investigate the model's extrapolation ability on unseen bands, we excluded one valence band and one conduction band across all 302 materials from training. Remarkably, as shown in Supplementary Fig. 9, the model achieves a high R^2=0.93 (MAE = 0.12 eV) on these two unseen bands. Therefore, our model can effectively predict GW corrections outside the training band



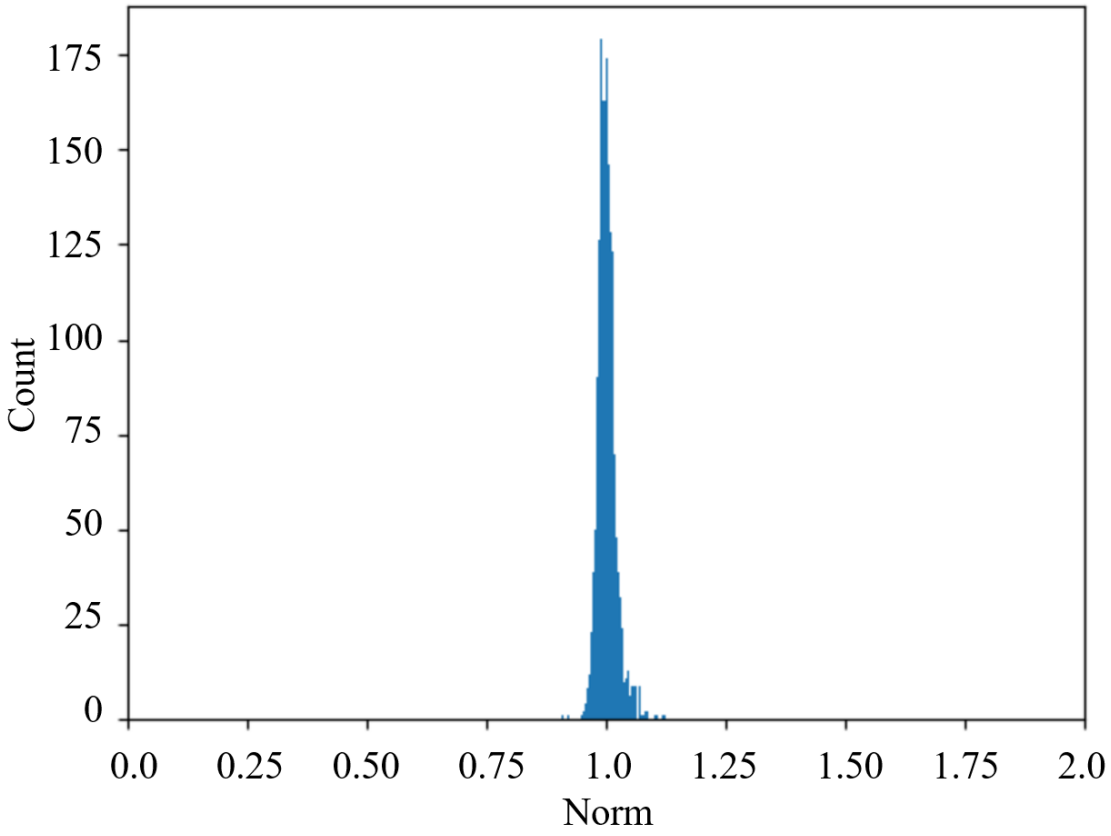
Supplementary Figure 9: Parity plot comparing the exact calculated values (x-axis) to the ML predicted values (y-axis) of the GW correction for reserved conduction and valence bands (1801 states in total), which are excluded from training across 302 materials. The R^2 for the test set are 0.93 (MAE = 0.12 eV). Source data are provided as a Source Data file.

Finally, in Fig. S2(f), we explore the importance of each input to the supervised NN in predicting GW corrections. Training is halted at 20,000 epochs before complete minimization of the loss function. This decision is based on the fact that the superstates and charge density for each individual state remain constant, then occasionally, the exclusion of KS states from the training can disrupt the entire process. Employing 20,000 training sets closely approximates the performance of the optimal model, making it suitable for comparative analysis.

Supplementary Note 3–Preservation of Physical Quantities

Preservation of realistic physical quantities is crucial to estimate the performance of our physical regression model and justify the whole workflow. In our training process, we define the total loss function with two components: i) the mean squared error (MSE) between the KS state before and after passing through the VAE, and ii) the KL divergence. The MSE component encourages each generated element of the KS states data to closely match the original states, which helps to preserve not only the pattern but also the normalization of the KS wavefunction after reconstruction. As Supplementary Fig. 10 shows, we use the well-trained VAE to generate KS states over the test dataset and present a histogram of their normalization. This demonstrates

that our approach effectively maintains the normalization of the KS wavefunction.



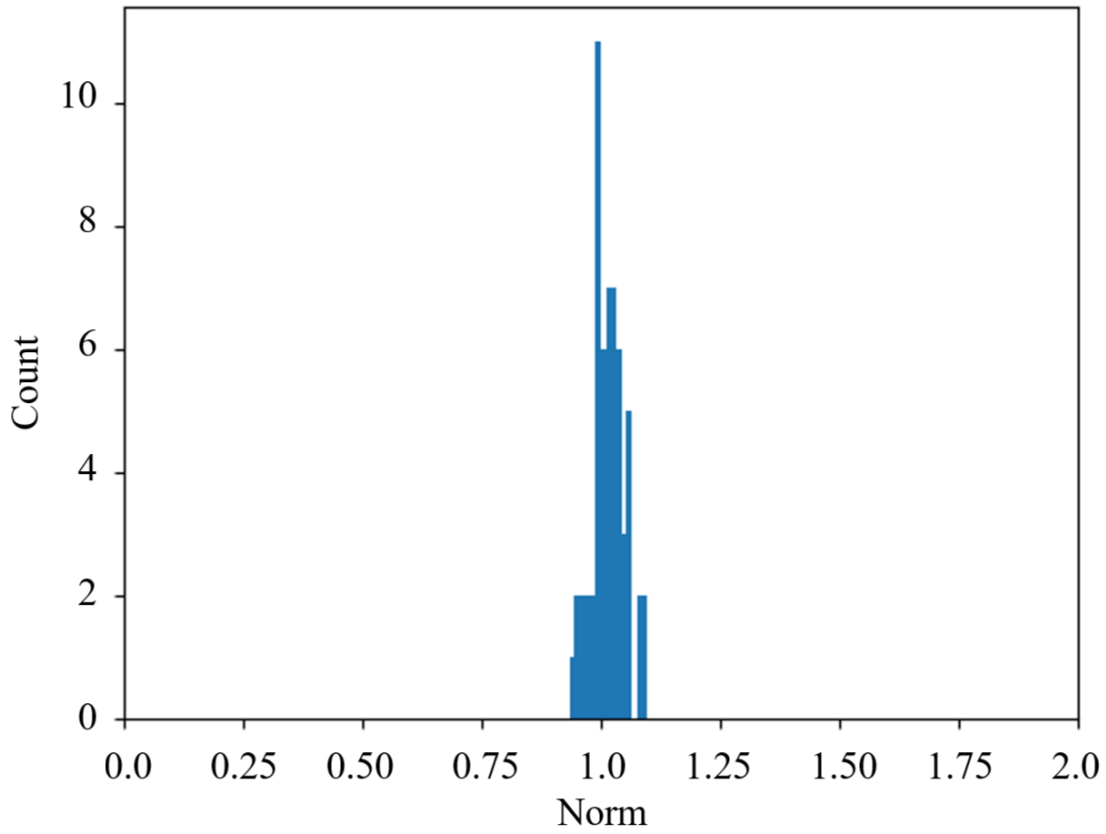
Supplementary Figure 10: The histogram of the norm of 1559 KS states reconstructed by VAE over test set. The mean absolute percentage error of reconstructed KS norm is 1%. Source data are provided as a Source Data file.

Secondly, we investigate if the KS states generated by the VAE-KS can reproduce the charge density on its own. We pass all occupied states of a material to VAE-KS and obtain the reconstructed KS state $\phi'_{nk}(r)$, then all reconstructed occupied states are summed to form the reconstructed charge density as follow:

$$\rho'(r) = \sum_{n,k}^{\text{occ}} |\phi'_{nk}(r)| \tag{9}$$

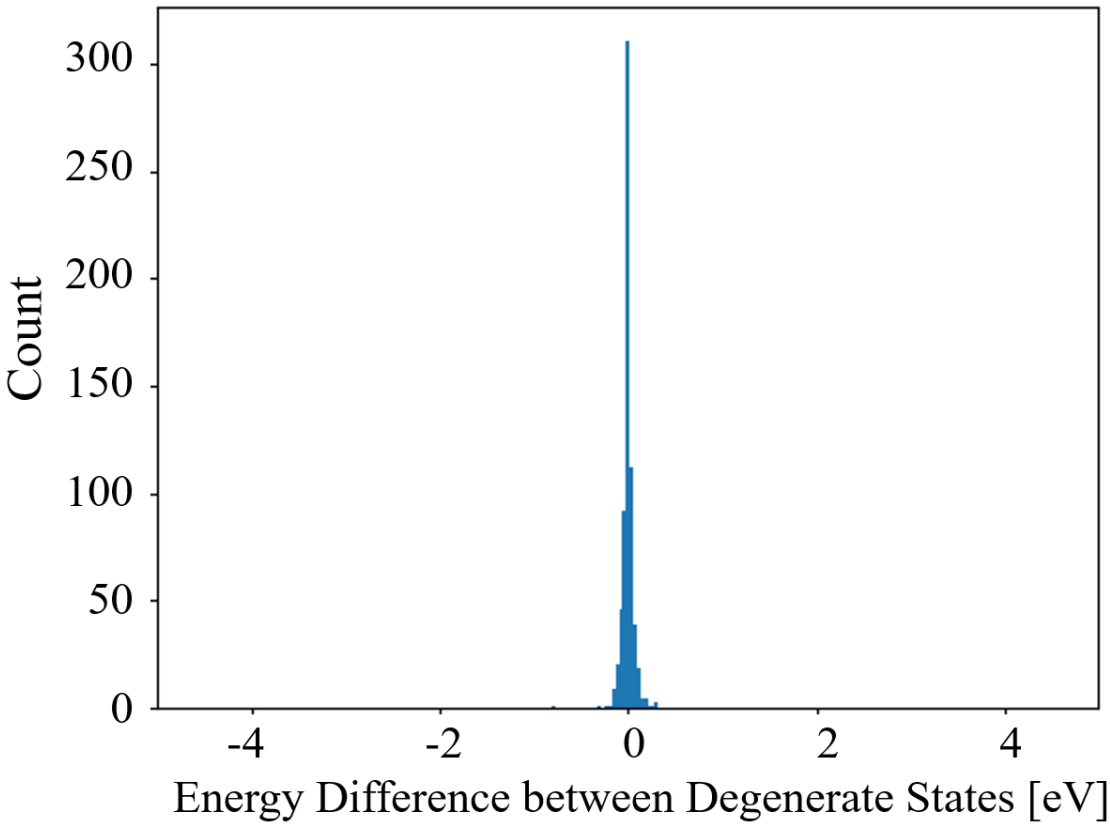
Intriguingly, even the VAE-KS performance of $R^2 = 0.99$ on the KS states in test set is 0.92, the reconstructed charge density by summing all occupied states across materials in the test set achieves a high $R^2 = 0.99$ compared against to the original charge density (We believe this might be attributed to cancellation of error). Therefore, the VAE, which is only trained with individual KS states, can also perform well in reproducing the ground state density.

Furthermore, We also check if the total charge is preserved before and after VAE. Supplementary Fig. 11 displays a histogram of the ratio of the charge number reconstructed by VAE-charge to the original charge density across the test/validation sets. The mean absolute percentage error of reconstructed KS norm is 3%. This demonstrates that the VAE-charge can effectively preserve the meaningful physical quantity, charge number, even when reconstructed from a low-dimensional latent space.



Supplementary Figure 11: The histogram of the norm of charge density of 61 materials reconstructed by VAE over test set. The mean absolute percentage error of reconstructed KS norm is 3%. Source data are provided as a Source Data file.

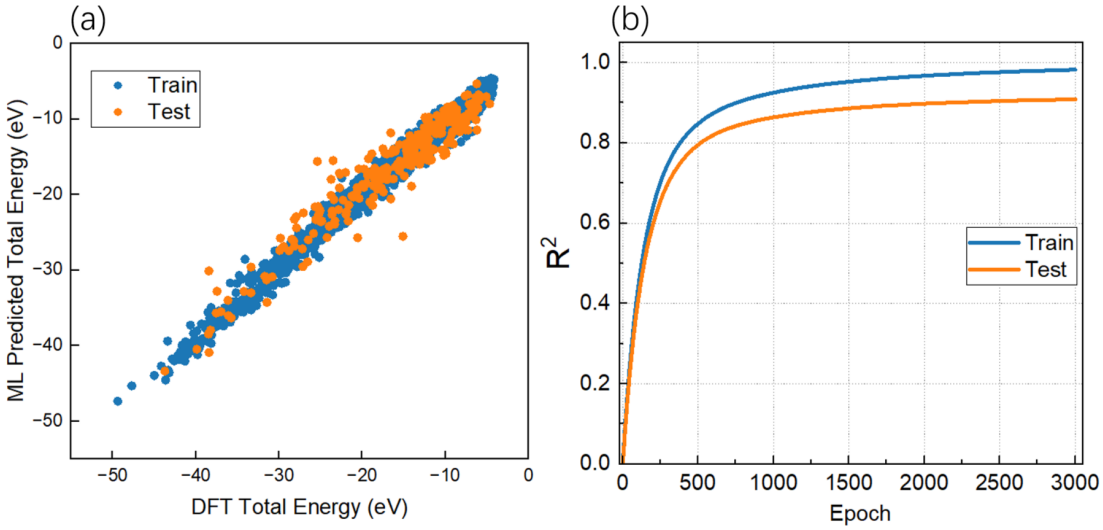
Lastly, to verify that the degeneracies at high-symmetry points are properly respected, we identified 663 pairs of degenerate states (a total of 1326 states) with threshold of 0.1 meV across 302 materials. We then used the trained model to predict their GW energies, resulting in a MAE of 0.11 eV, consistent with the model's performance on the test set. As shown in the Supplementary Fig. 12, we plotted the histogram of the energy difference $E_1 - E_2$ between the two degenerate states predicted by our model. The mean energy difference is 0.04 eV ($R^2 = 0.99$), which is significantly smaller than the total MAE. Therefore, we conclude that the degeneracies at high-symmetry points are well preserved.



Supplementary Figure 12: The histogram of energy difference between 663 pair of degenerate states. The mean energy difference is 0.03 eV with a high $R^2 = 0.99$. Source data are provided as a Source Data file.

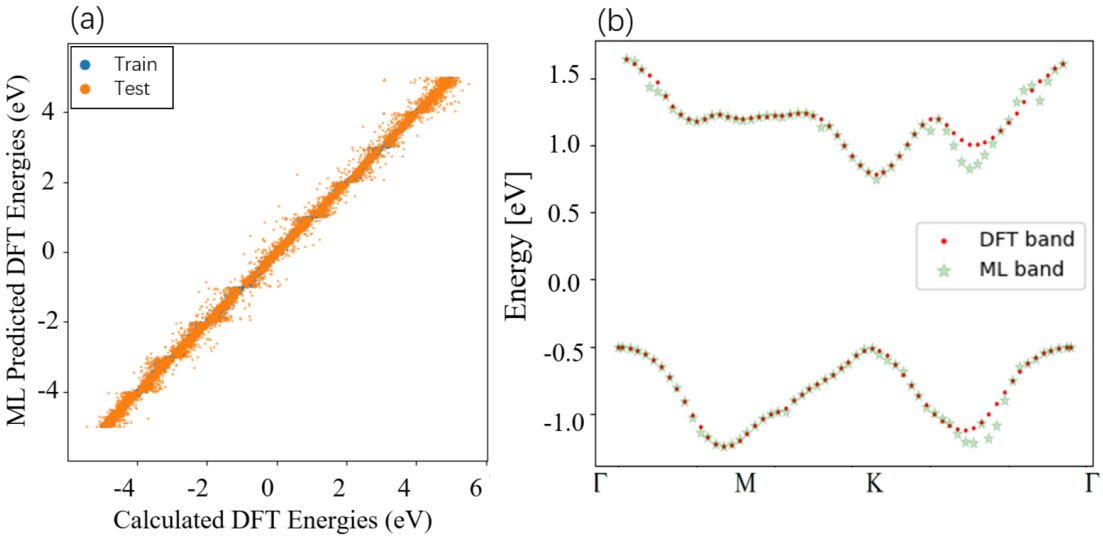
Supplementary Note 4–Application to Ground State Properties

Firstly, we explore how our VAE algorithm can generate a general low-dimensional representation of charge density, and how our model can be extended to other ground state physical quantities relevant to DFT. The prediction of the total ground state energy is a good benchmark of our model due to the explicit functional relation of $E[\rho]$. We train a simple regression model to learn the nonlinear mapping from $e_\theta(\rho)$ to the DFT total energy. Our dataset consists of 2670 2D materials, with each data point containing a DFT ground state charge density labeled with its corresponding total energy. Our dataset is randomly split into 90% for the training set and 10% for the validation/test set. As shown in the Supplementary Fig. 13(a), the MAE of the total energy is 1.5 eV for the test set and 0.7 eV for the training set, with a high a R^2 of 0.91 and 0.97, respectively. The training dynamics with respect to number of epochs are illustrated in Supplementary Fig. 13(b).



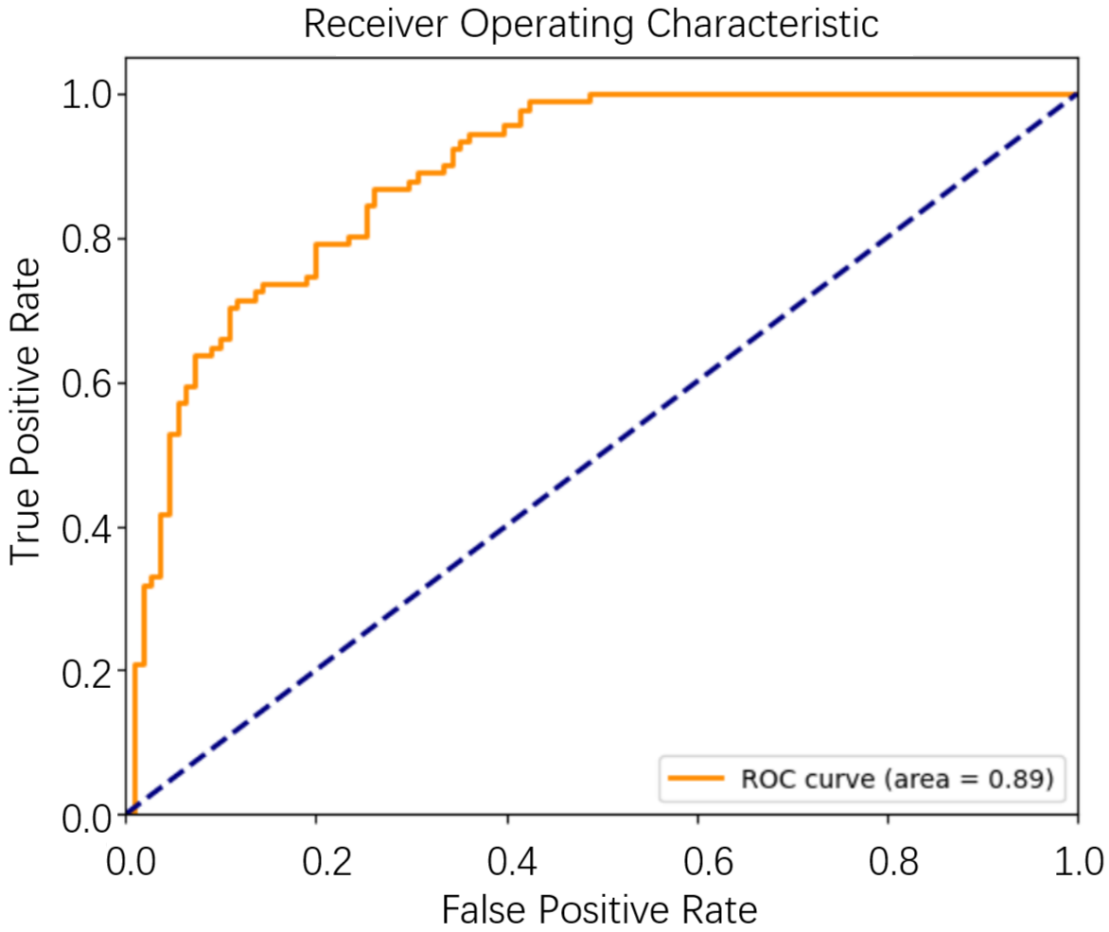
Supplementary Figure 13: (a) Parity plot comparing the exact calculated values (x-axis) to the ML predicted values (y-axis) of the DFT total energies for individual material. Blue (orange) dots represent training (test) sets. The training (test) set includes 2403 (267) states. The MAE for the training set and test set are 0.7 and 1.5 eV respectively. (b) Training dynamics of downstream NN for DFT total energies with respect to different number of epochs. The orange (blue) lines represent test (training) R^2. Source data are provided as a Source Data file.

In addition, unlike the GW self energy which has a closed form mapping relation, we still perform non-linear regression on top of the latent space of individual KS states to predict corresponding KS energies. As shown in Supplementary Fig. 14(a), we demonstrate that the direct prediction of the DFT eigenvalues can achieve fairly high accuracy. Our dataset consists of 16,760 states, with 40% allocated to the test set and 60% to the training set. The regression MAE over the test (train) set is 0.18(0.06) eV, as illustrated in the left figure below. As Supplementary Fig. 14(b) shows, the MoS_2 band structures for the first valence and conduction bands are displayed. Even we split the dataset in term of states instead of materials, our model still exhibits strong interpolation ability of DFT band predictions based on the KS state represebtations.



Supplementary Figure 14: (a) parity plot comparing the exact calculated values (x-axis) to the ML predicted values (y-axis) of the DFT energies for the individual state. Blue (orange) dots represent training (test) sets. The training set includes 10055 (6705) states The MAE for the training set and test set are 0.06 and 0.18 eV respectively. (b) ML predicted DFT band structures (green stars) and calculated PBE band structures (red dots) for monolayer MoS_2. Source data are provided as a Source Data file.

Lastly, in addition to performing regression, we also evaluate the classification capabilities of the VAE's low-dimensional representation of charge density. Specifically, we aim to classify whether a material is a metal or an insulator using this representation. To adapt the regression model to a binary classifier, we replace the loss function with cross-entropy loss for training the classification model. The dataset consists of 1006 materials, with 506 insulators and 500 conductors. We split the dataset into 80% for training and 20% for validation/testing. As Supplementary Fig. 15, the ROC AUC score achieves a high value of 0.89, which is comparable with previous ML model exclusively trained for metal/insulator classification[5].



Supplementary Figure 15: Receiver Operating Characteristic (ROC) curve for metal vs. insulator classification using VAE charge density representation: The model achieves a high area under the curve (AUC) value of 0.89, indicating strong performance in distinguishing between metal and insulator states. Source data are provided as a Source Data file.

Therefore, we believe our algorithm provides a very general low-dimensional representation, which can be widely applicable to the ground state density and other quantities relevant to DFT.

Supplementary Note 5–GW Calculation Details

In this study, we conduct *ab initio* mean-field calculations on 302 two-dimensional (2D) materials employing DFT in the Perdew–Burke–Ernzerhof (PBE) approximation[6]. Subsequently, GW calculations are performed, building on top of the DFT ground state wavefunctions. Both DFT and GW calculations are executed using GPAW software package [7, 8]. Here, the average vacuum region for the 302 2D materials is 14.78 Å(the average thickness of 302 materials is 3.41Å). There are 51 distinct chemical elements in our dateset.

For the supervised training for the GW prediction, we employ a cutoff energy of 400 eV for the planewave components of the DFT wavefunction. An 80 eV cutoff for planewave components

of the dielectric matrix and an average of 325 empty states is used to theoretically calculate 22,002 Σ^{nk} energies as supervised learning labels. GW corrections are calculated for 4 valence and 4 conduction states near Fermi level for each material. A uniform k-grid of $6 \times 6 \times 1$ is used for each material, which is comparable with previous ML models for GW[9, 3], which we use to benchmark the success of our model. Here, we note that our GW parameters are somewhat underconverged. They are, however, sufficient for the proof of the validation of the VAE representation for electronic structures, which is the main goal of this work.

Supplementary Note 6–Pseudobands Approximation

Constructing a ML model for GW prediction, which is both physics-informed and interpretable, presents an additional challenge. The diagonal term of the GW self-energy Σ^{nk} depends not only on $|n\mathbf{k}\rangle$, but also on all high energy empty bands (Eq. 1), which should all be fed to the neural network model. As a result, despite the considerable dimension reduction of each individual states through the VAE, the mapping from $\mathbb{R}^{N_{\text{train}} \times [(N+1)(p+1)]}$ to the self-energy space $\mathbb{R}^{N_{\text{train}}}$ is still challenging due to the substantial number of states N required for the GW calculation, i.e. $N_{\text{train}} \ll (N+1) \times (p+1)$, where p+1 represents the collective input of the latent space of KS states and its DFT energy.

In previous (non-ML) GW calculations, one way of compressing the empty states if through the pseudobands technique [10, 11, 12, 13], in which the empty bands are divided into one low-energy protected subspace and a series of pseudoband blocks, each of which combines bands with comparable energies. The energies of each block are then replaced by a single average energy, and for each $\mathbf{k}$ point, all $|n\mathbf{k}\rangle$ states in the block S are replaced by $\sum_n^S |n\mathbf{k}\rangle$. In this way, each pseudobands block is replaced by a nearly flat band, and at each $\mathbf{k}$ point, we have an effective, unnormalized wave function. We refer to this com this band is known as a pseudoband. We note that this process is not feature selection, as it has been shown that GW energies can be faithfully reconstructed by passing the pseudoband states and the average energies to the usual formulae for the GW self-energy. The information about the empty state dispersion relations near the Fermi surface lost in the averaging process is expected to be compensated by including the charge density into our model, which is also needed in standard Hybertsen-Louie Generalized Plasmon Pole (HL-GPP) approaches to GW[14]. In this section, we briefly discuss why this seemingly crude approximation works.

The expression of the zero-frequency dielectric matrix is

$$\chi_{\mathbf{G}\mathbf{G}'}(\mathbf{q}, \omega = 0) = \sum_{\mathbf{k}} \sum_{n}^{\text{occ}} \sum_{n'}^{\text{emp}} M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) M^*_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}') \frac{2}{E_{n\mathbf{k}+\mathbf{q}} - E_{n'\mathbf{k}}}, \tag{10}$$

where

$$M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) = \langle n\mathbf{k} + \mathbf{q} | \mathrm{e}^{\mathrm{i}(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}} | n'\mathbf{k} \rangle \,. \tag{11}$$

When pseudobands is used, the high-energy terms in $\chi_{\mathbf{G}\mathbf{G}'}(\mathbf{q}, \omega = 0)$ are therefore replaced by (P.B. means pseudobands)

$$\chi^{\text{P.B. terms}}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}, \omega = 0) = \sum_{\mathbf{k}} \sum_{S}^{\text{P.B. blocks}} \frac{2}{E_{n\mathbf{k}+\mathbf{q}} - \bar{E}_S} \sum_{n_1', n_2'}^{S} \sum_{n}^{\text{occ}} M_{nn_1'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) M^*_{nn_2'}(\mathbf{k}, \mathbf{q}, \mathbf{G}'). \tag{12}$$

To justify the pseudobands technique, it is sufficient to show that the unwanted $n_1' \neq n_2'$ cross terms that do not appear in the definition of χ vanish after the summations over $\mathbf{k}$ and n.

Because the energies of states in each pseudobands block are comparable, in each pseudobands block, the states share the same set of predominant $\mathbf{G}$ vectors. Thus, the states in pseudobands block S can be written as

$$\langle \mathbf{r} | n'\mathbf{k} \rangle = \frac{1}{\sqrt{V}} \sum_{i}^{N} c_{n'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) \mathrm{e}^{\mathrm{i}(\mathbf{k}+\mathbf{G}^{(i)}_{S\mathbf{k}})\cdot\mathbf{r}}, \quad \sum_{i}^{N} c_{n_1'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) c^*_{n_2'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) = \delta_{n_1'n_2'}, \tag{13}$$

and therefore

$$M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) = \sum_{i}^{N} c_{n'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) c^*_{n\mathbf{k}+\mathbf{q}}(\mathbf{G} + \mathbf{G}^{(i)}_{S\mathbf{k}}). \tag{14}$$

The pseudobands approximation of the diagonal $\chi_{\mathbf{GG}}$ components of the polarizability is proportional to

$$\begin{aligned}
&\sum_{n}^{\text{occ}} M_{nn_1'}(\mathbf{k},\mathbf{q},\mathbf{G}) M^*_{nn_2'}(\mathbf{k},\mathbf{q},\mathbf{G}) \\
&= \sum_{i,j}^{N} c_{n_1'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) c^*_{n_2'\mathbf{k}}(\mathbf{G}^{(j)}_{S\mathbf{k}}) \sum_{n}^{\text{occ}} c^*_{n\mathbf{k}+\mathbf{q}}(\mathbf{G}+\mathbf{G}^{(i)}_{S\mathbf{k}}) c_{n\mathbf{k}+\mathbf{q}}(\mathbf{G}+\mathbf{G}^{(j)}_{S\mathbf{k}}) \\
&\propto \sum_{i,j}^{N} c_{n_1'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) c^*_{n_2'\mathbf{k}}(\mathbf{G}^{(j)}_{S\mathbf{k}}) \delta_{ij} = \delta_{n_1'n_2'}.
\end{aligned} \tag{15}$$

In the third step, we argue that the summation over the occupied states leads to an approximate, non-normalized orthogonal relation, because the subspace spanned by the dominant $\mathbf{G}$ components of the occupied states is expected to largely overlap with the subspace of the occupied states, and therefore, although for $i = j$ terms, the summation over occupied n is usually far less than one, the fast oscillation of $i \neq j$ terms when n changes quickly brings the summation to zero. Therefore, the unwanted non-diagonal terms in (12) can be ignored when $\mathbf{G} = \mathbf{G}'$.

For the $\mathbf{G} \neq \mathbf{G}'$ terms we similarly have

$$\begin{aligned}
&\sum_{n}^{\text{occ}} M_{nn_1'}(\mathbf{k},\mathbf{q},\mathbf{G}) M^*_{nn_2'}(\mathbf{k},\mathbf{q},\mathbf{G}') \\
&= \sum_{i,j}^{N} c_{n_1'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) c^*_{n_2'\mathbf{k}}(\mathbf{G}^{(j)}_{S\mathbf{k}}) \sum_{n}^{\text{occ}} c^*_{n\mathbf{k}+\mathbf{q}}(\mathbf{G}+\mathbf{G}^{(i)}_{S\mathbf{k}}) c_{n\mathbf{k}+\mathbf{q}}(\mathbf{G}'+\mathbf{G}^{(j)}_{S\mathbf{k}}) \\
&\propto \sum_{i,j}^{N} c_{n_1'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}) c^*_{n_2'\mathbf{k}}(\mathbf{G}^{(i)}_{S\mathbf{k}}+\mathbf{G}-\mathbf{G}').
\end{aligned} \tag{16}$$

Numerical experiments reveal that the unwanted $n_1' \neq n_2'$ terms in the $\mathbf{G} \neq \mathbf{G}'$ components in the above equation do not cancel each other; the summation over $\mathbf{k}$ however could eliminate these terms. We note that the $1/(E_\text{v} - E_\text{c})$ factor in (12) is approximately a constant as $\mathbf{k}$ varies, and therefore the summation over $\mathbf{k}$ in (12) contains the following factor

$$\sum_{\mathbf{k}} c_{n_1'\mathbf{k}}(\mathbf{G}^{(i)}) c^*_{n_2'\mathbf{k}}(\mathbf{G}^{(j)}) \propto \sum_{\mathbf{k}} \mathrm{e}^{\mathrm{i}\theta_{n_1'\mathbf{k}} - \mathrm{i}\theta_{n_2'\mathbf{k}}} \overset{N_\mathbf{k}\to\infty}{\longrightarrow} \delta_{n_1'n_2'}, \tag{17}$$

where $\mathrm{e}^{\mathrm{i}\theta_{n\mathbf{k}}}$ is the random global phase factor introduced in the diagonalization process, which changes much more rapidly than the $\mathbf{G}$-dependence of $c_{n\mathbf{k}}$ as $\mathbf{k}$ runs over the Brillouin zone sampling, and therefore the summation in the equation above is proportional to the summation of the highly oscillating $\mathrm{e}^{\mathrm{i}(\theta_{n_1'\mathbf{k}} - \theta_{n_2'\mathbf{k}})}$ factor, which vanishes when $n_1' \neq n_2'$. In conclusion, the validity of the pseudobands technique for high-energy bands is well-justified for the whole $\chi_{\mathbf{GG}'}(\mathbf{q}, \omega = 0)$ because of the summation over $\mathbf{k}$ and/or n.

The analysis of pseudobands in `sigma` is more complicated because in the Coulomb hole part of the GPP self-energy

$$\begin{aligned}
\langle n\mathbf{k}\,|\Sigma_\text{CH}(E)|\,n'\mathbf{k}\rangle = \frac{1}{2} \sum_{n''} \sum_{\mathbf{q}\mathbf{G}\mathbf{G}'} & M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G}) M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G}') \\
& \times \frac{\Omega^2_{\mathbf{GG}'}(\mathbf{q})\,(1 - \mathrm{i}\tan\phi_{\mathbf{GG}'}(\mathbf{q}))}{\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})\,(E - E_{n''\mathbf{k}-\mathbf{q}} - \tilde{\omega}_{\mathbf{GG}'}(\mathbf{q}))} v\,(\mathbf{q}+\mathbf{G}'),
\end{aligned} \tag{18}$$

where $\Omega_{\mathbf{GG}'}$, $\tilde{\omega}_{\mathbf{GG}'}$ and $\phi_{\mathbf{GG}'}$ are quantities calculated from $\chi_{\mathbf{GG}'}(\mathbf{q}, \omega = 0)$ and the momentum space ground state density $\rho(\mathbf{G})$. The right hand side contains complicated dependence on $\mathbf{q}$, $\mathbf{G}$ and $\mathbf{G}'$. However, note that when n'' is high and is within one of the pseudobands blocks, typically the magnitudes of $\mathbf{G}$ and $\mathbf{G}'$ are large enough to dominate $v(\mathbf{q}+\mathbf{G}) = 4\pi/|\mathbf{q}+\mathbf{G}|^2$, and therefore the rapid variation of the random phase factor in $|n''\mathbf{k}-\mathbf{q}\rangle$ compared with the relatively slow variance of second line of (18) as $\mathbf{q}$ runs over the Brillouin zone sampling again

provides us with the opportunity to justify using pseudobands here: we have

$$\langle n\mathbf{k}|\Sigma_{\mathrm{CH}}^{\text{high energy bands}}(E)|n'\mathbf{k}\rangle \propto \sum_{\mathbf{G},\mathbf{G}'} \sum_{n''}^{\text{high energy bands}} \sum_{\mathbf{q}} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G}) M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G}'), \tag{19}$$

and by applying the arguments in (17), the pseudobands estimation of the high-energy terms in Σ_{CH} is

$$\begin{aligned}\langle n\mathbf{k}|\Sigma_{\mathrm{CH}}^{\text{P.B. terms}}(E)|n'\mathbf{k}\rangle &\propto \sum_{\mathbf{G},\mathbf{G}'} \sum_{S}^{\text{P.B. blocks}} \sum_{n_1'',n_2''}^{S} \sum_{\mathbf{q}} M^*_{n_1''n}(\mathbf{k},-\mathbf{q},-\mathbf{G}) M_{n_2''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G}') \\ &\propto \sum_{\mathbf{q}} |n_1''\mathbf{k}-\mathbf{q}\rangle \ \langle n_2''\mathbf{k}-\mathbf{q}| \overset{N_{\mathbf{q}}\to\infty}{\propto} \delta_{n_1''n_2''},\end{aligned} \tag{20}$$

and the unwanted $n_1'' \neq n_2''$ cross terms that do not appear in (19) again vanish.

The aforementioned fact that the random phase factor in DFT diagonalization ensures the pseudobands technique can be further exploited by *intentionally* inserting random phase factors before $|n\mathbf{k}\rangle$ states for a given $\mathbf{k}$ point and replacing a pseudobands block by several flat unnormalized bands, instead of just one such band; under such a scheme the size of pseudobands blocks can be drastically increased, further improving the speed of the *GW* methodology. Moreover, an empirical observation is that the valence bands can be pseudo-ized as well, and the protected subspace can be very small, without any real damage to accuracy [13].

The above discussion shows that there exists a relatively simple and smooth mapping from the protected bands, ground state density and pseudobands to the *GW* energies, which should be learnable for modern neural networks. Specifically, we note that adding pseudobands into the input data is *not* manual feature engineering, as the mapping from the protected bands, ground state density and pseudobands to the *GW* energies is mathematically established. There are however several further simplifications needed for a feasible machine learning model. Even after the pseudobands procedure, there are still hundred of (protected and pseudo) bands; in our preprocessing procedure, we tentatively eliminate the protected subspace and pseudo-ize all bands, and also reduce the number of pseudobands subspaces to 3 for each material. We also average pseudo-states in one pseudoband in the dimension of $\mathbf{k}$ to further reduce the size of the input data, hoping that random phase factor cancellation mechanisms similar to what is outlined above will ensure the validity of this procedure. Since the ground state electron density in principle contains all information of the material, as shown by the foundation of DFT [15], having the (compressed) ground state electron density as a part of the input to the model compensates for the information loss.

References

[1] Arunkumar Chitteth Rajan et al. "Machine-learning-assisted accurate band gap predictions of functionalized MXene". In: *Chemistry of Materials* 30.12 (2018), pp. 4031–4038.

[2] Joohwi Lee et al. "Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques". In: *Physical Review B* 93.11 (2016), p. 115104.

[3] Nikolaj Rørbæk Knøsgaard and Kristian Sommer Thygesen. "Representing individual electronic states for machine learning GW band structures of 2D materials". In: *Nature Communications* 13.1 (2022), p. 468.

[4] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[5] Alexandru B Georgescu et al. "Database, features, and machine learning model to identify thermally driven metal–insulator transition compounds". In: *Chemistry of Materials* 33.14 (2021), pp. 5591–5605.

[6] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized Gradient Approximation Made Simple". In: *Phys. Rev. Lett.* 77 (18 1996), pp. 3865–3868.

[7] Jens Jørgen Mortensen et al. “GPAW: An open Python package for electronic structure calculations”. In: *The Journal of Chemical Physics* 160.9 (2024).

[8] Ask Hjorth Larsen et al. “The atomic simulation environment—a Python library for working with atoms”. In: *Journal of Physics: Condensed Matter* 29.27 (2017), p. 273002.

[9] Asbjørn Rasmussen, Thorsten Deilmann, and Kristian S Thygesen. “Towards fully automated GW band structure calculations: What we can learn from 60.000 self-energy evaluations”. In: *npj Computational Materials* 7.1 (2021), p. 22.

[10] Mauro Del Ben et al. “Large-scale GW calculations on pre-exascale HPC systems”. In: *Computer Physics Communications* 235 (2019), pp. 187–195.

[11] Weiwei Gao et al. “Speeding up GW calculations to meet the challenge of large scale quasiparticle predictions”. In: *Scientific reports* 6.1 (2016), p. 36849.

[12] Weiwei Gao et al. “Quasiparticle energies and optical excitations of 3C-SiC divacancy from *GW* and *GW* plus Bethe-Salpeter equation calculations”. In: *Physical Review Materials* 6.3 (2022), p. 036201.

[13] Aaron R. Altman, Sudipta Kundu, and Felipe H. da Jornada. “Mixed Stochastic-Deterministic Approach for Many-Body Perturbation Theory Calculations”. In: *Phys. Rev. Lett.* 132 (8 2024), p. 086401.

[14] Mark S. Hybertsen and Steven G. Louie. “Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies”. In: *Phys. Rev. B* 34 (8 1986), pp. 5390–5413.

[15] Pierre Hohenberg and Walter Kohn. “Inhomogeneous electron gas”. In: *Physical review* 136.3B (1964), B864.