nature portfolio

Peer Review File

Unsupervised Learning of Individual Kohn-Sham States: Interpretable Representations and Consequences for Downstream Predictions of Many-Body Effects

Corresponding Author: Professor Diana Qiu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In their paper, Hou and coworkers use machine learning techniques to demonstrate that Kohn-Sham wavefunctions obtained from first-principles density-functional theory (DFT) calculations can be dramatically compressed. Based on this discovery, they develop a machine learning approach to predict GW quasiparticle band structure of materials which include many-body corrections not captured by DFT. GW calculations are significantly more challenging than DFT calculation and therefore the ability to obtain GW band structures efficiently using machine learning techniques constitutes a significant advance. In light of these significant advances, I recommend the paper for publication in Nature Communications after the authors have addressed the following questions.

+ there are lots of typos in the manuscript – I recommend the authors carefully go over the manuscript and fix them.

+ why do the authors use the KS wavefunctions on a real space grid as input for the VAE? Why do they not work with the Fourier components?

+ in Eq. 1, a different mathematical symbol should be used for the normalization factor 1/n as n is already used as a band index

+ what happens when the VAE is not used in the GW prediction? How sensitive are the results to the degree of data compression? As the authors state the GW training data is quite small even compared to the NN input data even after the VAE compression. This makes me wonder how important the compression actually is.

+ when ML models for dielectric functions are discussed (line 155), it would appropriate to cite the recent paper by Zauchner et al. on “Accelerating GW calculations through machine-learnt dielectric matrix” (npj computational materials 2023).

+ in Eq. 2, the RHS is a function of the frequency w; however, in the discussion the authors refer to this quantity as single number – I suspect because the self energy is evaluated at the quasiparticle energy; in this case, w should be set to the quasiparticle energy in Eq. 2

+ sometimes nk is used as subscript and sometimes as a superscript, would be good to be consistent.

+ in the paragraph following Eq. 3, the subscripts nk are missing in the equation for the GW self energy

+ are the KS still normalized after compression and decompression by the VAE?

+ it is not clear to me if the VAE is only used as a compression tool or can it also predict wavefunctions; the papers says for example that MoS2 is not in the training set, but I suspect this only refers to the GW part of the paper – some clarification of this point would be useful.

+ Fig 2 a and b: axes should have labels with units

+ Fig 2 e: it is hard to compare the quality of the ML band structure: it would be better to have the explicit GW band structure as lines (not just the results for a few kpoints);

+ Fig 2d) there are some large outliers ; can the authors comment on those?

+ Fig 3 a) and b): it would be good if the colour of the band structures in panel b would be the same as those of the corresponding paths in a). Also, the DFT energies in both panels appear to be very different: in a) they range from -2.2 eV to -1.6 eV while in b) they range from -1 eV to 0 eV.

Reviewer #2

(Remarks to the Author)
I found the manuscript by Hou et al., and in particular how they explore using a variational autoencoder (VAE) to create an interpretable representation of Kohn-Sham wavefunctions, intriguing and valuable. The analysis of the VAE latent space looks promising for advancing applications in the field. However, I have some reservations that I would like to address.

The requirement for a fixed input size significantly limits the applicability of the VAE and of the neural network to a class of similar materials. The use of a uniform real-space grid across all wavefunctions, regardless of the unit cell size and symmetry, overly simplifies the system and likely results in inadequate sampling if extended to general materials. Inserting an initial graph neural network layer would allow for inputs of variable sizes and better respect the symmetry of the system.

Similarly, the fixed output size constrains the NN utility. The model predicts GW corrections for only four valence and four conduction bands across a fixed 6x6x1 k-point grid. This one-size-fits-all approach does not accommodate the diverse energy windows and Brillouin zone sizes determined by different crystal symmetries, which likely explains why other studies, relying on larger material databases, have opted to predict only single values, like the band gap. This is of course a problem with no easy solution, but still would require some more thought.

I do not think it is correct to state that this model provides a "prediction of the quasiparticle band structure within the GW formalism". The model allows to predict n,k dependent GW corrections to the Kohn-Sham band structure, after having calculated the Kohn-Sham band structure. Is the VAE enabling the prediction of a "smooth, physically realistic" band structure or is it simply predicting that the GW corrections to a Kohn-Sham band structure are almost constant (justifying the common use of scissor operators to open up the gap)? It would be beneficial to know whether the network can also directly predict Kohn-Sham band structures. In particular, are degeneracies at symmetry points respected? Where are minima and maxima? etc. The model's approach of predicting near-constant corrections to the Kohn-Sham band structure reduces the output to essentially a single value, which is too simplistic and limits the depth of analysis that could be achieved.

A deeper statistical analysis of the results beyond the mean absolute error (MAE) is necessary. The presentation of a single band structure is not sufficiently representative of the model's performance. It is crucial to analyze how the GW correction varies across different k-points and bands and whether the model accurately respects symmetries at high-symmetry points. What is the MAPE? What is the maximum error and for which kind of materials?

More details are needed regarding the training process, such as how the training progresses with increases in the number of data points and epochs. This information is crucial for understanding learning dynamics.

Given these concerns, I am currently unable to recommend this manuscript for publication in Nature Communications as it stands. Enhancements in the model flexibility and a more comprehensive evaluation of its predictive capabilities and training dynamics, are needed before it can meet the publication standards.

Reviewer #3

(Remarks to the Author)
In this work, the authors present a useful methodology for learning single-particle Kohn Sham wavefunctions (KS) from density functional theory (DFT). They use an unsupervised VAE encoding scheme to learn a low-dimensional latent space representation for the KS wavefunctions, which can then be used to decode accurate predictions for KS wavefunctions outside of the training data set. They also wrap a GW band structure prediction learning algorithm around their KS learning algorithm, and find good agreement between its predictions and first principles GW band structure predictions. The unsupervised nature of the KS learning algorithm is perhaps the most intriguing aspect, as the authors have mentioned that most other DFT learning algorithms require heavy-handed feature selection. I believe the authors make significant contributions and present their results in a convincing manner in this manuscript.

I have two comments to make regarding the manuscript, after which being addressed I believe it would be suitable for publication.

The first is that the authors' have focused, in this paper, on reproducing the single-particle KS wavefunctions. This is evident by looking at their cost function, which is a least-squares over single particle wavefunctions. The issue therein is that the KS wavefunctions themselves are not a physically relevant quantity in DFT, the electronic density is. It is unclear, then, whether the learning algorithm performs well in reproducing the ground state density properly. Furthermore, since most DFT XC functionals involve complex non-local integrals over the density, it is also unclear whether this algorithm provides any significant improvements in the ground state density properties relevant to DFT. I do understand that the authors focus in this manuscript is the reproduction of single-particle band structures, for which the learning algorithm may be suitable, but it seems warranted to discuss errors in densities and total ground state DFT energies relative to the first principles data. This should be easy to manage, since the authors must have already computed the first principles DFT energies and densities en route to the results they have in their manuscript.

Secondly, I think some discussion about potential applications to other post-HF corrections may be useful (maybe just a paragraph in the conclusion). For example, using the ML generated wavefunctions in quantum Monte Carlo (QMC). In particular, there is a methodology for developing low-energy Hamiltonians which relies on efficient representations of low-energy wavefunctions called Density Matrix Downfolding [https://arxiv.org/abs/1712.00477]. This method usually involves a lot of user supervision, especially in sampling and constructing low-energy wave functions, but it seems that the method shown by the authors in this manuscript may be able to provide a clean unsupervised way to do this, at least at the DFT level. Given that most QMC wave functions are built from DFT wavefunctions, the connection to the prior should be straightforwards. In general, some discussion on post-HF applications of the VAE wavefunctions would be useful, since many post-HF (QMC, CCSD, CASCI) involve DFT wave functions as a starting point.

Lastly, I recommend the authors rephrase statements such as "wiggles in band structures" to something more appropriate and precise. Similar kinds of statements are made throughout the manuscript.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have carefully and convincingly replied to all my questions and comments. I recommend that the paper is published.

Reviewer #2

(Remarks to the Author)
The authors replied convincingly to all questions and I like their modified model that accounts for symmetries. The revised manuscript includes new results, more analysis and technical information. It is therefore much clearer. I recommend to accept the manuscript for publication.

Reviewer #3

(Remarks to the Author)
The authors have addressed comments in my original comments thoroughly. I would recommend this manuscript for acceptance at this stage.

Reply to referee 1:

In their paper, Hou and coworkers use machine learning techniques to demonstrate that Kohn-Sham wavefunctions obtained from first-principles density-functional theory (DFT) calculations can be dramatically compressed. Based on this discovery, they develop a machine learning approach to predict GW quasiparticle band structure of materials which include many-body corrections not captured by DFT. GW calculations are significantly more challenging than DFT calculation and therefore the ability to obtain GW band structures efficiently using machine learning techniques constitutes a significant advance. In light of these **significant advances**, I recommend the paper for publication in Nature Communications after the authors have addressed the following questions.

We thank the referee for his or her positive evaluation and recommendation for publication pending clarifications of some questions. The referee's careful review and suggestions have allowed us to improve the work and strengthen our conclusion. Below, we address the referee's concerns point by point.

There are lots of typos in the manuscript – I recommend the authors carefully go over the manuscript and fix them.

We thank the referee for their careful review. We have revised the typos in the manuscript.

Why do the authors use the KS wavefunctions on a real space grid as input for the VAE? Why do they not work with the Fourier components?

We appreciate the reviewer's question regarding the use of real-space grid input for the VAE training instead of Fourier components. Firstly, following comments from Referee 2, we have revised our model to explicitly recognize **translational invariance, periodic boundary conditions, discrete rotational symmetry and unit cell size without any data augmentation.** The new model can work for both real space grids and Fourier components. However, we still want to explain the reason we prefer working in the real space instead of reciprocal space in our previous workflow.

In the DFT calculations, we fix the kinetic cutoff energy for the planewave basis across all materials. Consequently, the number of planewaves varies across different materials depending on the material's lattice constant. However, in our original model, although the CNN itself doesn't require a fixed input size, the output of the CNN must be fixed for the input of the next dense layer. Thus, we resized the wavefunction grid to a standard size (e.g. $30 \times 30 \times 40$), following conventions for training CNNs followed by fully connected NNs used for image processing. Thus, we preferred to use the real-space grid for image resizing because the KS state in real space is smoother, allowing for interpolation during resizing. The interpolation of G space, on the other hand, is less smooth, since higher G-vectors correspond to higher frequency oscillations in real space.

Despite some limitations of using simple cropping/resizing techniques, our old model still worked well for predicting GW self energies. This is because the expression for GW energies does not explicitly require knowledge of lattice constants and symmetry; only the components of the wavefunction come in explicitly. However, the fixed input size of the VAE by cropping/resizing could lead to potential information loss or substantial distortion. This posed a further challenge for generalizing our model to wider applications, such as recognition of defects in large supercells. Therefore, we have enhanced the model architecture by designing a special CNN layer combining rotational CNNs, circular padding, and a global pooling technique. These modifications allow the encoder to accept inputs of any size and also enable proper handling of periodic boundary conditions, translational invariance, and rotational symmetry. Consequently, both real space grids and reciprocal space grids (resizing is not required in the new model) can equivalently be used for training the VAE.

In Eq. 1, a different mathematical symbol should be used for the normalization factor 1/n as n is already used as a band index.

We now use T to represent the normalization factor for the training dataset.

What happens when the VAE is not used in the GW prediction? How sensitive are the results to the degree of data compression? As the authors state the GW training data is quite small even compared to the NN input data even after the VAE compression. This makes me wonder how important the compression actually is.

We appreciate the reviewer's question regarding the impact of data compression on the final GW predictions. High-dimensional data, such as wavefunctions and scattering vertex matrices, remain a bottleneck of ML in condensed matter physics (and is the topic of much recent work see Refs. [1-5]).

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 result 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 Fig. R1. Our results indicate that overfitting decreases with higher levels of data compression.

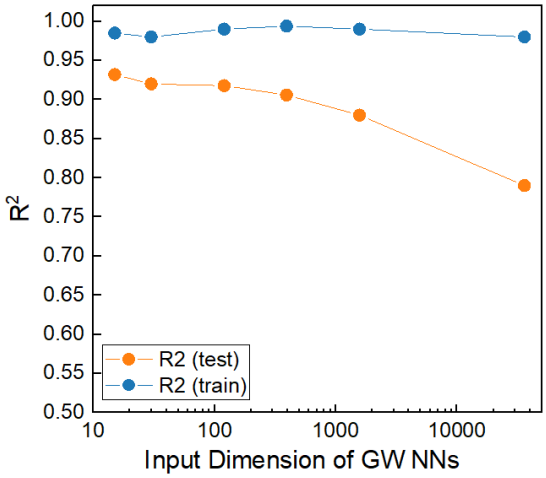

Figure R1. R^2 of GW prediction with respect to different input dimensions of KS states. The orange (blue) dots represent model performance on test (training) set.

iii) High training memory requirement. For example, the DFT wavefunction across 302 materials is, in our case, ~ 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.

We have added this information to the supplemental material (Supplementary Fig. 8).

When ML models for dielectric functions are discussed (line 155), it would appropriate to cite the recent paper by Zauchner et al. on "Accelerating GW calculations through machine-learnt dielectric matrix" (npj computational materials 2023).

We have cited this paper in our discussion on dielectric functions.

In Eq. 2, the RHS is a function of the frequency w; however, in the discussion the authors refer to this quantity as single number – I suspect because the self-energy is evaluated at the quasiparticle energy; in this case, w should be set to the quasiparticle energy in Eq. 2

The reviewer is correct. We have fixed Eq. 2 by setting the frequency ω to the quasiparticle energy $\varepsilon_{n,\vec{k}}$.

Sometimes nk is used as subscript and sometimes as a superscript, would be good to be consistent.

We have moved all nk index to subscript to ensure consistent notation throughout the manuscript.

In the paragraph following Eq. 3, the subscripts nk are missing in the equation for the GW self energy.

We have fixed missing subscript for the GW self-energy.

Are the KS still normalized after compression and decompression by the VAE?

We thank the reviewer for raising this important point. This is a crucial point for justifying our workflow, which aims to preserve physical information during training. 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 Fig. R2 shows, we use the 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.

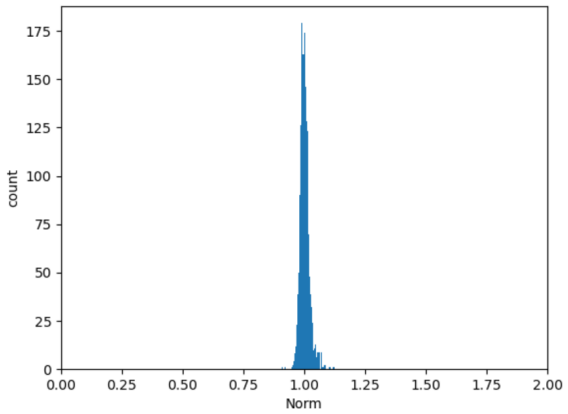

Figure R2. The histogram of the norm of KS states reconstructed by VAE over the test set. The mean absolute percentage error of reconstructed KS norm is 1%.

We have added this information to the supplemental material (Supplementary Fig. 10).

It is not clear to me if the VAE is only used as a compression tool or can it also predict wavefunctions; the papers says for example that MoS2 is not in the training set, but I suspect this only refers to the GW part of the paper – some clarification of this point would be useful.

First, we would like to clarify that MoS_2 is not in the training set for the VAE or the GW model. In the manuscript, to validate the VAE training, we withheld 10% of the dataset as a test set, achieving a high R^2 score of 0.92. We further exclude three monolayer TMD materials—MoS_2, WS_2, and CrS_2—from both the VAE and GW training datasets. We have revised our manuscript to emphasize this point and added more details of training dynamics to the SI

Regarding the generative power of the model, the key motivation for using VAE is that unlike other compression techniques that have been used in condensed matter physics (e.g. Singular

Value Decomposition (SVD), Principal Component Analysis (PCA), and Autoencoders (AE) Refs. [2-6]), the VAE can serve as both an information extractor and wavefunction generator.

In the manuscript, the generative ability is demonstrated for wavefunction interpolation in Fig. 3(b). Through regularization of the KL divergence, the VAE models the probability distribution of the KS wavefunction conditioned on a low-dimensional latent variable, resulting in a smooth latent space[6], which is suitable for downstream regression. This allows us to generate wavefunctions at points in k-space that have never been seen in the training set. Additionally, as noted by the reviewer, we visualize the latent space of the generated wavefunctions for the three TMDs, MoS_2, WS_2, and CrS_2 (Fig.3 (a)), which are never seen in the VAE training set. More generally, we hope our model possesses generative capabilities, which are applicable to more diverse scenarios. For example, if a low-dimensional latent space can accurately reconstruct the original wavefunction, we can potentially develop a simple deep neural network to learn the mapping from (crystal, n, k) to a corresponding point in the latent space. As a result, this enables a controllable and low-cost wavefunction generator or predictor for other physical observables, but the demonstration goes beyond the scope of the current work.

Fig 2 a and b: axes should have labels with units

We apologize for the missing units in Fig. 2 (a-b). All real space wavefunction squared moduli are presented in fractional coordinates based on the crystal lattice constants. We have added units to the figures and caption.

Fig 2 e: it is hard to compare the quality of the ML band structure: it would be better to have the explicit GW band structure as lines (not just the results for a few kpoints)

We appreciate the reviewer's suggestion to use calculated GW band structures as reference lines for visual clarity. One thing to note is that the GW calculation is explicitly performed at a small number of k-points (represented by the dots on the bandstructure). The lines in the GW bandstructure represent an interpolation of those explicitly calculated points. Below, we overlay the complete GW band structure onto Fig. 2(b) (Fig. R3).

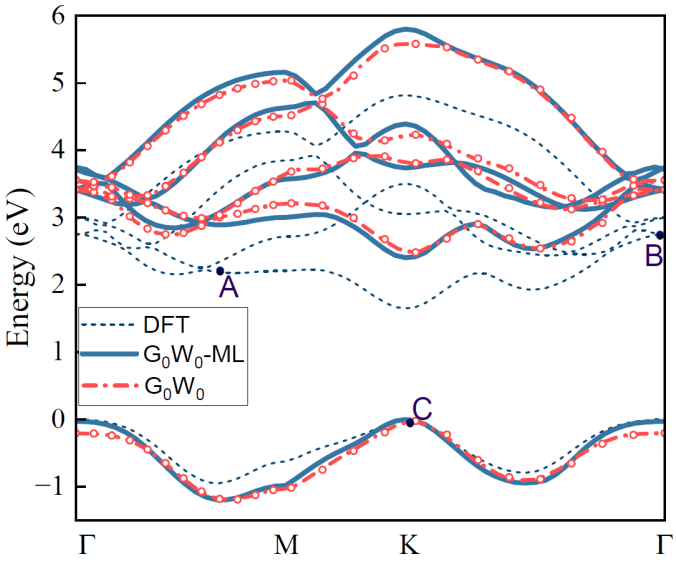

Figure R3. ML predicted GW band structures (blue solid curve) and calculated PBE band structures (blue dashed line) for monolayer MoS2. The red circles are the exactly calculated QP energies from GW self-energy. The red dashed lines are the interpolated GW band structures by GPAW.

We have replaced Fig. 2 (e) with the new bandstructure figure.

Fig 2d) there are some large outliers; can the authors comment on those?

We appreciate the reviewer's question regarding the large outliers in our regression predictions. To identify these outliers, we filtered out materials and states with predicted GW correction errors 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 is not quite meaningful given this situation. A possible explanation for this variability is that some GW corrections may not be fully converged, resulting in "noise" that is learnt by the model, leading to outliers.

Fig 3 a) and b): it would be good if the color of the band structures in panel b would be the same as those of the corresponding paths in a). Also, the DFT energies in both panels appear to be very different: in a) they range from -2.2 eV to -1.6 eV while in b) they range from -1 eV to 0 eV.

To enhance consistency, we have modified the color of the band structures in Fig. 3(b) to match the latent path shown in Fig. 3(a). In Fig. 3(b), we aligned the valence band maximum (VBM) to 0 eV to facilitate comparison of the band structures of the three TMD materials. In contrast, Fig. 3(a) shows the original results from our DFT calculations. To ensure consistency between Fig. 3(a) and (b), we have also adjusted the energy range of the color bar in Fig. 3(a) so that the Fermi level aligns with 0 eV.

Reply to referee 2:

I found the manuscript by Hou et al., and in particular how they explore using a variational autoencoder (VAE) to create an interpretable representation of Kohn-Sham wavefunctions, intriguing and valuable. The analysis of the VAE latent space looks promising for advancing applications in the field. However, I have some reservations that I would like to address.

We thank the referee for his or her careful review of our work. The reviewer posed substantial valuable comments and questions regarding the model architecture, potential applications, limitations, additional benchmarks, statistical analysis, etc., which have allowed us to significantly improve our work. Below, we address the reviewer's comments point by point.

The requirement for a fixed input size significantly limits the applicability of the VAE and of the neural network to a class of similar materials. The use of a uniform real-space grid across all wavefunctions, regardless of the unit cell size and symmetry, overly simplifies the system and likely results in inadequate sampling if extended to general materials. Inserting an initial graph neural network layer would allow for inputs of variable sizes and better respect the symmetry of the system.

We appreciate the review's thoughtful comment and suggestion that using a fixed uniform (equidistant) real-space grid across all wavefunctions might constrain our model to the limited set of materials. Indeed, as originally implemented, we would expect our model to only be effective for training and test sets of materials with similar lattice constants, and it would be difficult to generalize, for instance, to large supercells. This insight has helped us significantly enhance our model to extend its applicability to a broader range of materials and pave a practical way for wider applications beyond downstream many-body predictions.

While the referee suggested inserting a graph layer might be helpful for respecting the symmetry of the crystal and handling variable input size, we decided to take a different approach. We understand that graph neural networks (GNN) might increase the learning efficiency for some models purely based on simple crystal geometry due to its inherent graph structure. However, our approach is motivated by the fact that the KS wavefunctions are scalar fields across the space regardless of the original geometry of the crystal, and it is, thus, essentially grid (matrix)-based data, similar to an image. Therefore, introducing a graph network may not be the most consistent solution for our approach.

The main bottlenecks faced by our previous model based on a simply rigid and uniform grid are: i) lack of preservation of translational invariance, ii) lack of preservation of discrete rotational symmetry, iii) improper handling of periodic boundary conditions, and iv) lost information of unit cell size and symmetry. To account for this, we take inspiration from developments in ML for image processing and make the following modifications to our model:

(i) We implement a circular CNN to treat periodic boundary conditions
(ii) We add a global average pooling (GAP) layer to treat translational symmetry and unit cells with different sizes in real space.
(iii) We implement a rotational CNN to preserve the invariance of selecting different lattice vectors of the unit cell.

(iv) We have added extensive additional benchmarking, showing that our modified VAE can recognize translational and rotational symmetry, as well as supercells.

The grid size (resolution), which carries the wavefunction in real space within a primitive unit cell, depends on both the crystal lattice and the kinetic energy cutoff of the plane wave basis in the DFT calculations. In our old VAE model, even though the number of trainable parameters in the CNN layers remained constant regardless of the size of the input grid, the fully connected dense layer always required the same input size from the output of the CNN layer. To address this, a commonly used preprocessing technique is cropping/resizing before passing input to a CNN (e.g., a fixed 30×30×40 grid in our old model). However, as the referee mentioned, this straightforward method does not adequately account for the variations in unit cell size, kinetic cutoff energies, periodic boundary condition (PBC) and translational symmetry, which limits wider application.

To enhance our model and eliminate the dependence on fixed input sizes, we adopt the global average pooling[7] (GAP) technique. Unlike traditional methods such as cropping and resizing, the GAP layer enables CNNs to accept inputs of varying sizes and produce a fixed-length output, which is crucial for the fully connected layers in the encoder. More importantly, by aggregating the spatial information of the feature maps, the global pooling layer provides translational invariance with respect to the unphysical degree of freedom that comes from choosing the origin point of the unit cell. Apart from translation invariance and unit cell size, periodic boundary conditions (PBC) are another challenge of dealing with Bloch states in real space within a unit cell. To address this, we swap the original CNN layers with circular CNNs[8], which include PBC padding in the convolutional layers. Lastly, to achieve invariance to the choice of lattice vectors, we implemented a discrete rotational CNN along lattice constant directions and max out the combined feature maps for any input. Since a single layer of rotational CNN is sufficient to ensure coordination lattice invariance, we integrated it into the first layer of the encoder. These invariances are crucial for downstream many-body predictions, as the physical observables should not vary with the choice of unit cell. Finally, to ensure that the VAE-generated wavefunction's size aligns with the input, allowing us to define a proper loss function, we include an adaptive layer as the final layer in our decoder.

Here, we will mathematically demonstrate why the VAE latent space from such an encoder can effectively represent the KS state of a specific material with unique unit cell. Suppose we want to learn the representation of a 2D KS state:

$$|\varphi_n(x, y, c)|I_{x,y\in(-\infty,+\infty)}$$

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:

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

, where the 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 CNN-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, we firstly 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:

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

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

$$Relu = \frac{x+|x|}{2}$$

Here, 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 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:

$$\tilde{a}_n(C') = \sum_X^{T_x}\sum_Y^{T_y} \tilde{O}_n(X,Y,C';t_x,t_y) = \sum_X^{T_x}\sum_Y^{T_y} \tilde{O}_n(X,Y,C')$$

, where the unphysical degree of freedom of t_x and t_y are summed out with X and Y. As a result, we achieve physical invariance to arbitrary translations through this global channel aggregation, and the consequent output vector $\vec{\tilde{a}}_n \in \mathbb{R}^{C_{out}}$ 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 a single unit cell with minor modifications. Suppose we consider the KS state of a $M \times N$ super cell:

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

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

$$\begin{aligned} \tilde{A}_{M,N,n}(C') &= \sum_i^N \sum_j^M \sum_X^{T_x}\sum_Y^{T_y} \tilde{O}_n(X+i\times T_x, Y+j\times T_y, C') \\ &= (M\times N)\times \sum_X^{T_x}\sum_Y^{T_y} \tilde{O}_n(X,Y,C') \end{aligned}$$

$$= (M \times N) \times \tilde{a}_n(C')$$

, where the second line comes from the integral invariance of a periodic function. Intriguingly, the representation of an $(M \times N)$ super cell $\tilde{A}_{M,N,n}(C')$ are always equal to $(M \times N)$ times that of a unit cell $\tilde{A}_{M=1,N=1,n}(C')$. Since our new 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:

$$\vec{z}_{M,N,n}(\mathrm{C}') = \frac{1}{NT_x \times MT_y}\tilde{A}_{M,N,n}(\mathrm{C}') = \frac{1}{T_x \times T_y}\tilde{a}_n(\mathrm{C}') = \frac{1}{T_x \times T_y}\tilde{A}_{1,1,n}(\mathrm{C}') = \vec{z}_{1,1,n}(C')$$

As a result, the latent space $\vec{z}_n = f(\vec{z}_{1,1,n}(C'))$ (f is the following dense layer) remains the same for any supercell size, and it can represent a complete KS state density $|\varphi_n(x,y,c)|I_{x,y\in(-\infty,+\infty)}$ 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 continuously complete periodicity of a specific KS state will generate the same $\vec{z}_{1,1,n}(C')$. In this approach, the symmetry is implicitly preserved even if symmetry labels are never included in the training data, and the grid-based data can be properly used as the input layer if a continuous periodicity is ensured. Here, $\vec{z}_{1,1,n}(C')$ can be interpreted as a vector of the "average charge density" from different nonlinear global feature maps of a state.

In summary, the new model has several advantageous properties to generate effective representations: i) $\vec{z}_n$ is always smooth due to the smooth nature of neural network, ii) $\vec{z}_n$ can handle translational invariance and PBC, iii) $\vec{z}_n$ can handle any cell size and symmetry , iv) Even only trained with unit cell data, $\vec{z}_n$ can be extended to large system beyond the unit cell, which provides future opportunities to study systems like defects and moiré bilayers.

We have modified our model, accordingly, as shown below and in the revised Fig. 1(a) (Fig. R4), integrating the circular CNN layer, global average pooling layer, and rotational CNNs (Fig. 1(b)).

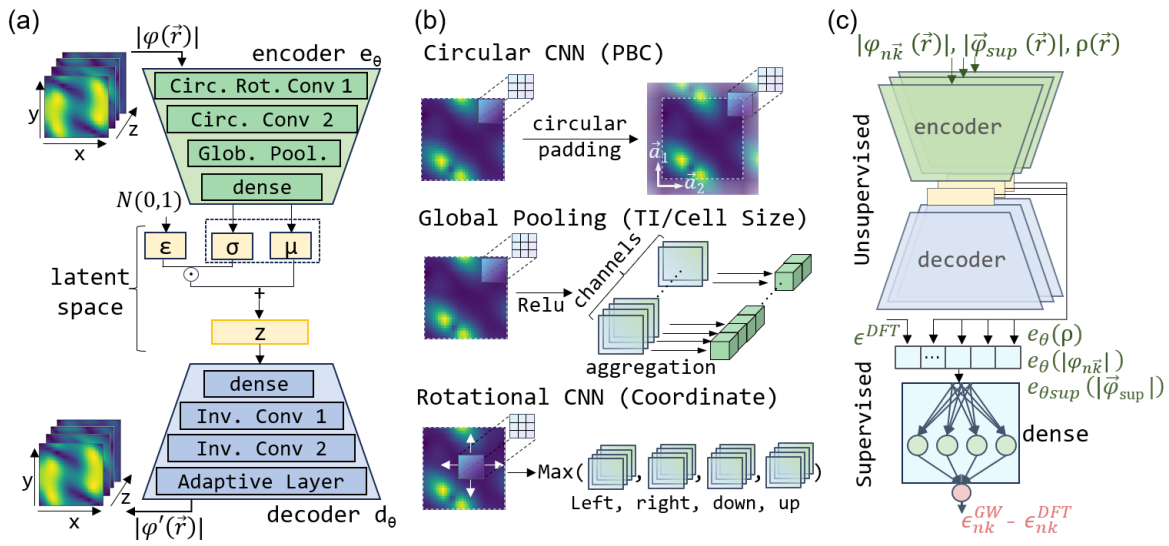

Figure R4. (a) Schematic of the VAE. The encoder (green trapezoid) consists of a circular rotational CNN, a circular CNN, a global pooling layer and one flattened dense layer, which are explicitly presented in (b). The encoder maps a real space wavefunction to a latent space vector of variational mean $\vec{\mu}$ and variance $\vec{\sigma}$. Z is the sampled latent vector, drawn from a variational Gaussian distribution using a "reparameterization trick". The decoder (blue trapezoid) has symmetric NN structures. The latent space serves as an information bottleneck for the VAE as its dimensions are only $10^3 - 10^4$ smaller than the input and output (represented by the colormaps of the wavefunction in real space). (b) Circular CNN layer includes the circular padding techniques based on the periodic boundary condition. The width of the circular padding equals to the size of CNN kernel so that CNN output size has the same periodicity as the input. Global pooling layers outputs the average value of each channel from CNN feature map. Rotational CNN layers scan the input along four directions and outputs the max of (left, right, up, down) to the next layer. TI denotes translational invariance, PBC denotes periodic boundary condition. The detail of model parameters are listed in the SI. (c) Schematic of the overall semi-supervised learning model, including both the unsupervised VAE and supervised dense neural networks (NN). The VAE inputs are the KS wavefunction modulus $|\varphi_{n\vec{k}}(\vec{r})|$, all super states $|\varphi_{sup}(\vec{r})|$ and charge density $\rho(\vec{r})$ in real space. The input layer of the supervised dense NN is comprised of DFT energies, denoted as ε^{DFT}, along with low dimensional effective representations of $\varphi_{n\vec{k}}(\vec{r})$, $\varphi_{sup}(\vec{r})$ and $\rho(\vec{r})$ denoted as $e_\theta(\varphi_{n\vec{k}}(\vec{r}))$, $e_{\theta sup}(\varphi_{sup}(\vec{r}))$ and $e_{\theta\rho}(\rho(\vec{r}))$ respectively. These representations are encoded within the VAE latent space (yellow square) through an encoder with parameters θ, θsup and θ_ρ, which are unsupervisedly trained for all KS wavefunctions, super states and charge density.

We have replaced Fig.1 with the updated ML model scheme.

After training, the new VAE model achieves a high R^2 of 0.91 on the test set and 0.93 on the training set, and the downstream GW prediction based on the latent space of our new VAE remains at the same prediction accuracy as before. Therefore, all our previous conclusions still hold true.

Next, we demonstrate additional generalizability of our model beyond the previous approach. In the first row of Fig. R5 (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 Fig. R5(a-f) displays the VAE latent space and reconstructed KS wavefunction (insets), obtained by inputting the corresponding shifted wavefunction into our new model. The latent space and generated wavefunction remains entirely invariant to any sliding of the unit cell. The third row demonstrates the latent space and generated wavefunction using our old model, which are sensitive to the selection of the unit cell. We see that the modified VAE significantly enhances our model's ability to deal with systems of different unit cell size, cutoff energies, PBC and translational invariance.

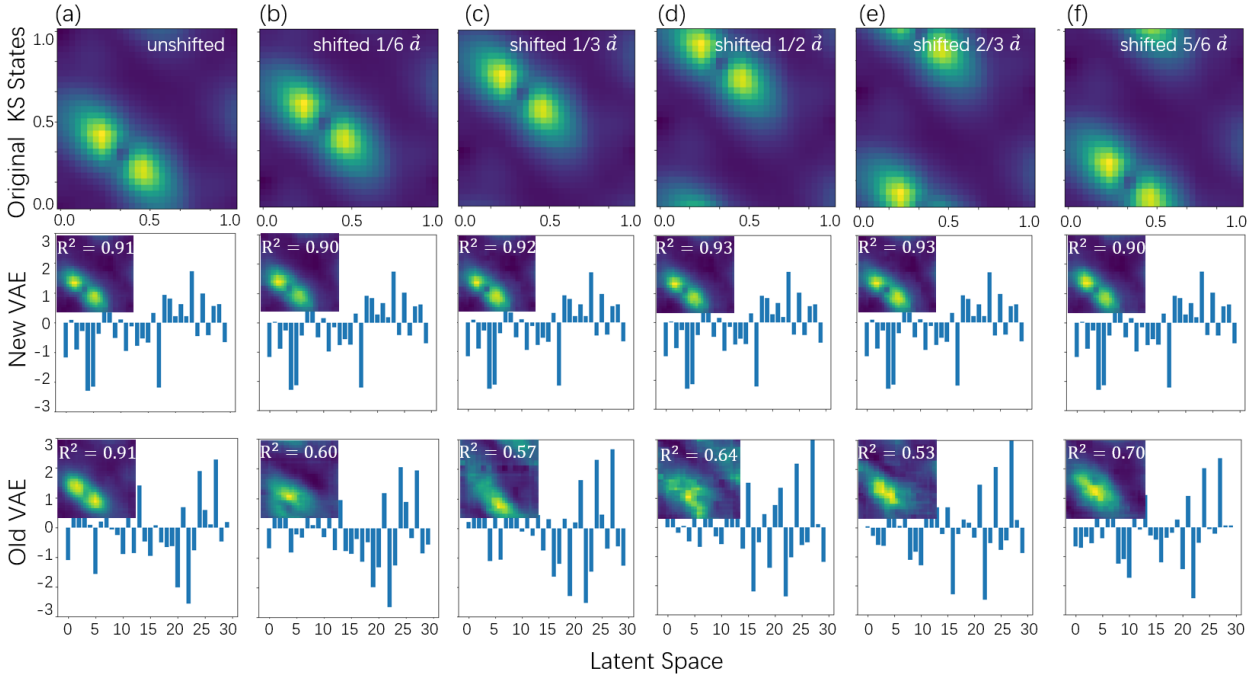

Figure R5. the first row of (a-f) shows a randomly picked KS state shifted along y-direction with 0, 1/6, 1/3, 1/2, 2/3 and 5/6 $\vec{a}$ in real space. The second row of (a-f) illustrates the latent space of related shifted KS states respectively. The insets are the corresponding reconstructed wavefunction, 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 unit cell, which preserve the translational invariance for the downstream physical prediction. The third row of (a-f) is latent space and reconstructed KS states by old VAE, which are presented for comparison.

Next, our new model is independent of the size of the input in real space, and the encoder output is invariant to the repetition of the periodicity. As a result, it can be easily extended to large super cells as Figure R6 illustrates:

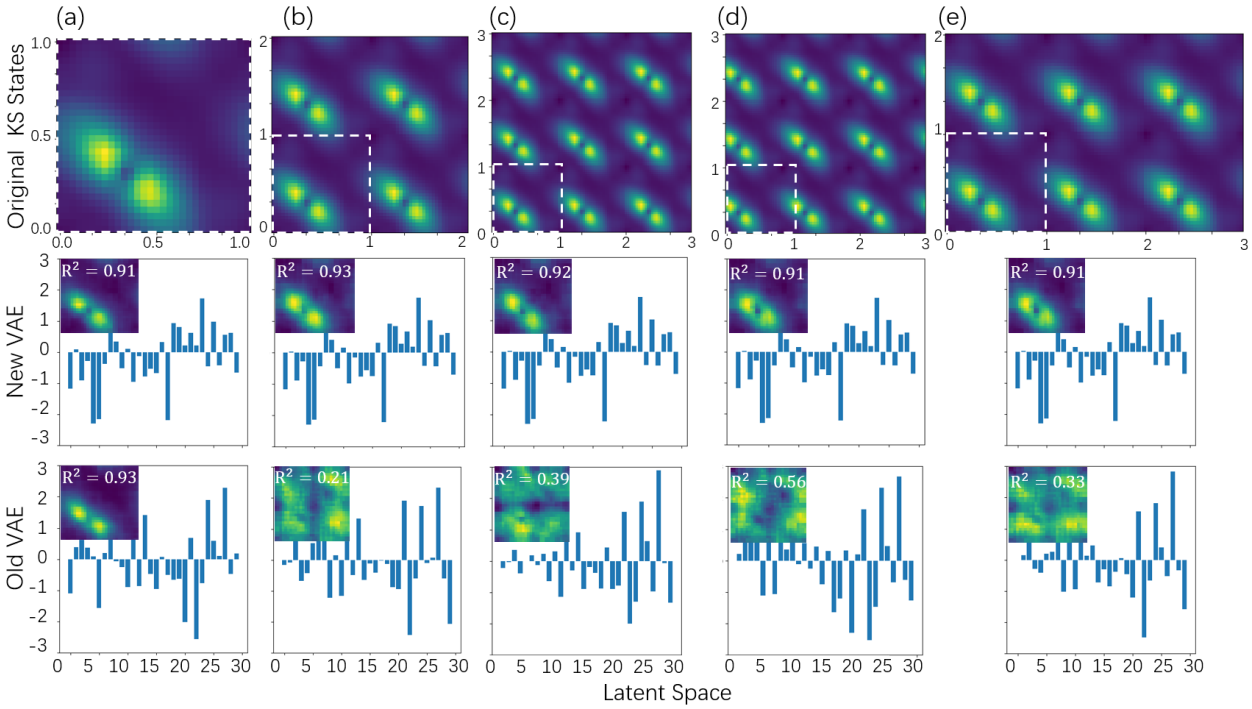

Figure R6. the first row of (a-e) illustrate the 1×1, 2×2, 3×3, 3×3 (shifted) and 2×3 super cell of a randomly picked KS state. The second row of (a-e) show the latent space of related KS states from different super cell respectively. The insets are the corresponding reconstructed wavefunction, generated by decoding the latent space of KS state from different super cell respectively. 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 old VAE, which are presented for comparison.

Lastly, to investigate how the new model handles the degree of freedom of selecting the coordinate basis, we chose four combinations of lattice vectors $(\vec{a}_1, \vec{a}_2)$, $(\vec{a}_2, -\vec{a}_1)$, $(-\vec{a}_1, -\vec{a}_2)$, and $(\vec{a}_1, -\vec{a}_2)$ to form the crystal unit cell. These combinations correspond to 90, 180, and 270-degree rotations of the wavefunction in fractional coordinates, as shown in the first row of Fig. R7 (a-d). The second (third) row of Fig. R7 (a-d) illustrates the latent space and reconstructed wavefunction corresponding to the different choices of unit cell basis through our new (old) VAE. As expected, the new VAE maintains the original pattern, while the old VAE is sensitive to the choice of coordinate basis.

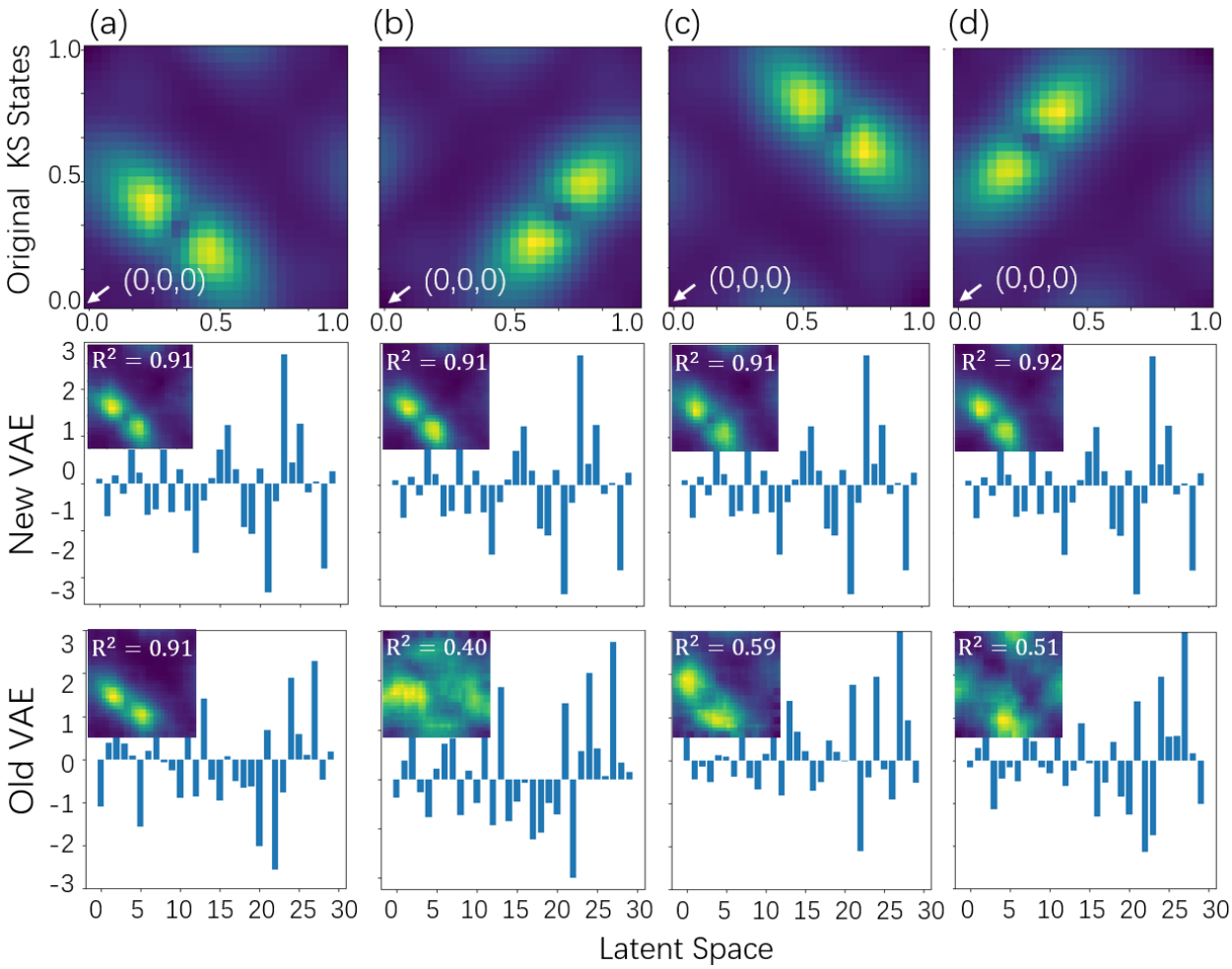

Figure R7. the first row of (a-d) illustrate a randomly picked KS state represented in a unit cell with bases of lattice vectors $(\vec{a}_1, \vec{a}_2)$, $(\vec{a}_1, -\vec{a}_2)$, $(-\vec{a}_2, \vec{a}_1)$, and $(\vec{a}_2, -\vec{a}_1)$. The second row of (a-d) illustrates the latent space of related rotated KS states respectively. The insets are the corresponding reconstructed wavefunction, generated by decoding the latent space of rotated KS state respectively. The latent space (reconstructed wavefunction) is invariant to the choice of the lattice vectors, which are suitable for the downstream physical prediction. The third row of (a-d) are reconstructed KS states by old VAE, which are presented for comparison.

We have added this information to the supplemental material (Supplementary Fig. 1-3).

Similarly, the fixed output size constrains the NN utility. The model predicts GW corrections for only four valence and four conduction bands across a fixed 6x6x1 k-point grid. This one-size-fits-all approach does not accommodate the diverse energy windows and Brillouin zone sizes determined by different crystal symmetries, which likely explains why other studies, relying on larger material databases, have opted to predict only single values, like the band gap. This is of course a problem with no easy solution, but still would require some more thought.

We thank the referee for their comments on the output of our model. However, we respectfully disagree with the points raised by referee and believe there may be some misunderstanding

regarding the workflow of our model, particularly concerning the output and training process. We apologize for any lack of clarity in our original statement and have revised the manuscript to make it more explicit. Here, we need to make two crucial clarifications in response to the referee's concerns:

i) Four valence and four conduction bands across a fixed 6x6x1 k-point grid are the states used for training the NN model, **not** the output size of the model. These states are chosen for convenience in generating training data and are irrelevant to the structure of the NN model. In addition, the choice of training set is not limited to any specific grid size either.

ii) Due to its strong extrapolation ability, even though the $\hat{f}_{NN}$ model is trained with only a few k-points and bands, 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. 3(c), training the model at a specific k-point enables it to predict other nearby k-points accurately, demonstrating the model's generalizability across arbitrary k-points. Furthermore, to investigate the model's extrapolation ability to unseen bands, we excluded one valence band and one conduction band across all 302 materials from training. Remarkably, as shown in Fig. R8, the model achieves a high $R^2 = 0.93$ (MAE = 0.12 eV) for these two unseen bands. Therefore, our model can effectively predict GW corrections outside the training bands. This is why we can obtain the full GW band structure (see Fig. 2(e)), despite the fact that the training data only includes a few points in the BZ and bands.

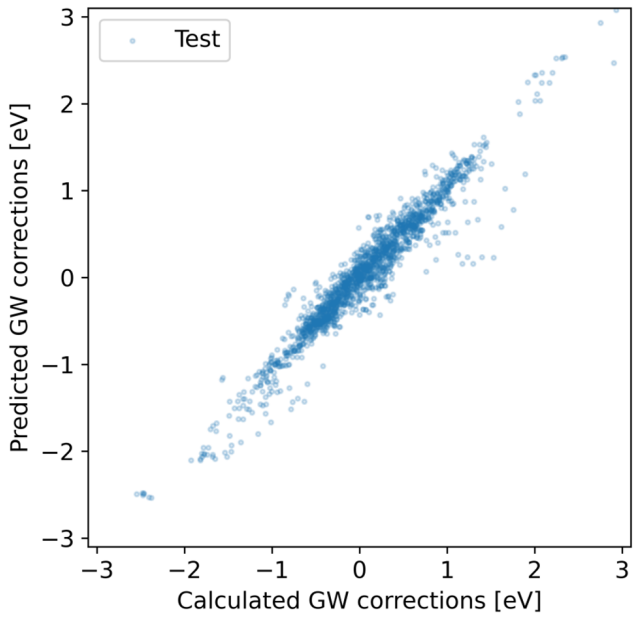

Figure R8. Parity plot comparing the exact calculated values (x-axis) to the ML predicted values (y-axis) of the GW correction for reserved conduction and valence bands, which are excluded from training across 302 materials. The R^2 for the test set are 0.93 (MAE = 0.12 eV)

In summary**,** our model is not fixed to a specific k-grid or a set number of valence or conduction bands, it can predict $\Sigma_{n\vec{k}}^{G_0W_0}$ for any arbitrary KS state from only the latent space of that specific KS state.

We have added this information to the supplemental material (Supplementary Fig. 9).

I do not think it is correct to state that this model provides a "prediction of the quasiparticle band structure within the GW formalism". The model allows to predict n,k dependent GW corrections to the Kohn-Sham band structure, after having calculated the Kohn-Sham band structure. Is the VAE enabling the prediction of a "smooth, physically realistic" band structure or is it simply predicting that the GW corrections to a Kohn-Sham band structure are almost constant (justifying the common use of scissor operators to open up the gap)? It would be beneficial to know whether the network can also directly predict Kohn-Sham band structures. In particular, are degeneracies at symmetry points respected? Where are minima and maxima? etc. The model's approach of predicting near-constant corrections to the Kohn-Sham band structure reduces the output to essentially a single value, which is too simplistic and limits the depth of analysis that could be achieved.

We appreciate the referee's feedback regarding the statement that our model provides a 'prediction of the quasiparticle band structure within the GW formalism.' However, we stand firmly by our statement and have revised our manuscript to include additional data supporting our claim. Below, we address the referee's concerns point by point.

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[9,10] or after carefully tailoring features, were able to predict a k-point-dependent GW correction that lacked the smoothness of a physical bandstructure [2]. 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 (and remarkably), unlike previous work that carefully tailored feature selection for the GW problem specifically, our representation learning is not designed to target GW, and it can be easily generalized in the future to any downstream method relying on DFT wavefunctions.

Secondly, we address the question of whether the GW correction can be captured by a scissor shift. 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. Fig. R9 (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, Fig. R9(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 exceeds 0.1 eV as well. Similarly, Fig. R9(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 (or scissor shift), and the statement that we are “predicting near-constant corrections to the Kohn-Sham eigenvalues” is incorrect.

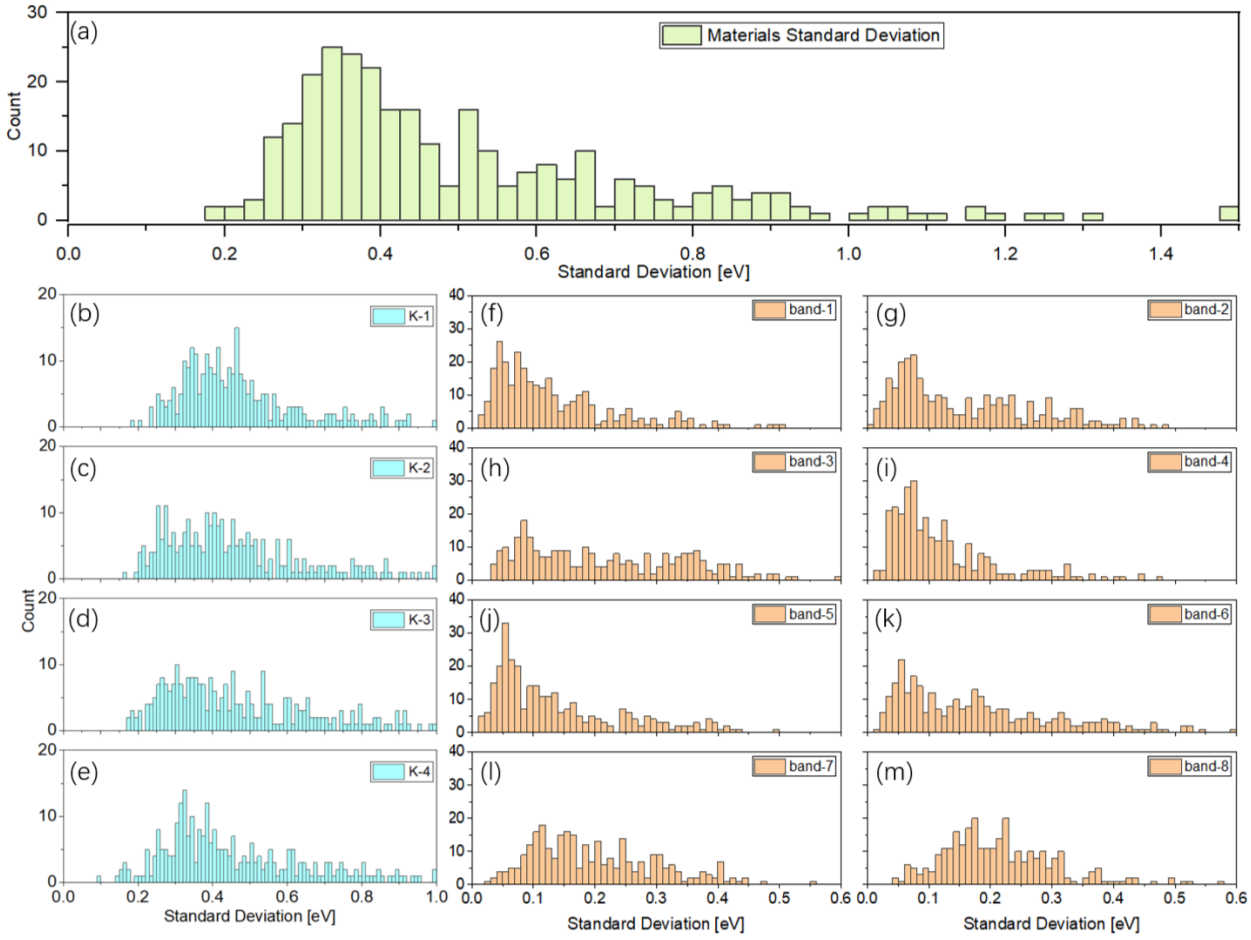

Figure R9 (a) the histogram of standard deviations of GW corrections from all bands and k-points across all 302 materials. (b-f) the histogram of standard deviation of GW corrections from all bands given the same k-point across all 302 materials. (f-m) the histogram of standard deviation of GW corrections from all K-points given the same band across 302 materials.

Thirdly, the referee suggests that we should directly predict DFT band structures from representations of KS states. However, it is important to note that KS states and energies result from the diagonalization of the DFT Hamiltonian, unlike the GW self energy which has a closed form mapping relation. Thus, the reconstruction of the original eigenvalues solely from eigenstates is not mathematically guaranteed based on the current design of our model, and we would argue falls beyond the scope of this work. Despite this uncertainty, we perform non-linear regression on top of the latent space of individual KS states to predict corresponding KS energies and demonstrate that we can actually predict the DFT eigenvalues to fairly high accuracy, which is shown in Fig. R10 (a). Our dataset consists of 16,000 states, with 20% allocated to the test set and 80% to the training set. The regression MAE over the test (train) set is 0.18(0.06) eV, as illustrated in the Fig. R10(a). As Fig. R10 (b) shows, the MoS2 band structures for the first valence and conduction bands are displayed in the right panel. We note that due to the lack of closed form

mapping, the VAE representation is only effective for interpolation, not extrapolation, and our bandstructure is produced by including 50% of the states of MoS_2 in the training set.

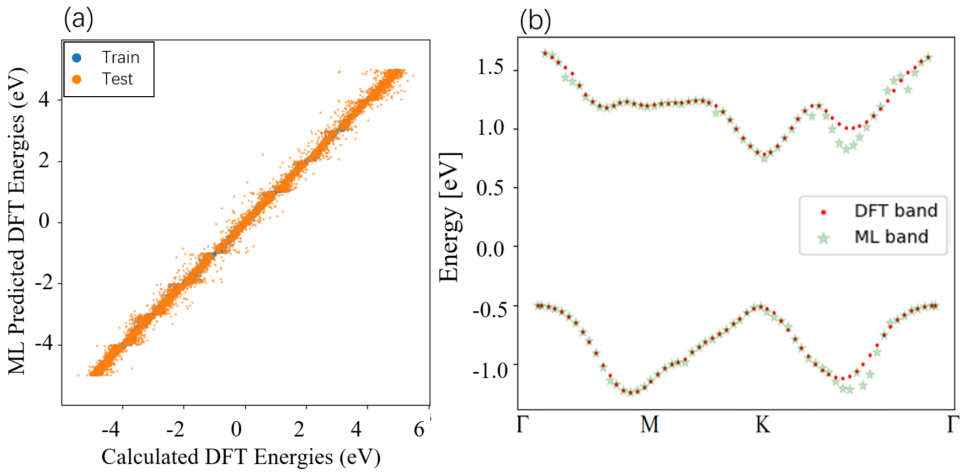

Figure R10 (a) parity plot comparing the exact calculated values (x-axis) to the ML predicted values (y-axis) of the DFT energies for the individual state. Blue (orange) dots represent training (test) sets. The MAE for the training set and test set are 0.06 and 0.18 eV respectively. (b) ML predicted DFT band structures (green stars) and calculated PBE band structures (red dots) for monolayer MoS_2.

Finally, 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 Fig. R11, 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 0.03 eV ($R^2 = 0.99$), which is significantly smaller than the total MAE. Therefore, we conclude that the degeneracies are reasonably well preserved. We note that the slight degeneracy lifting is consistent with the result of standard GW calculations due to the arbitrary rotation of the degenerate subspace. In standard GW codes, degeneracies are imposed manually by averaging the GW correction of degenerate states. The calculated self energy Sigma, which is the function that we are trying to learn, can break degeneracy..

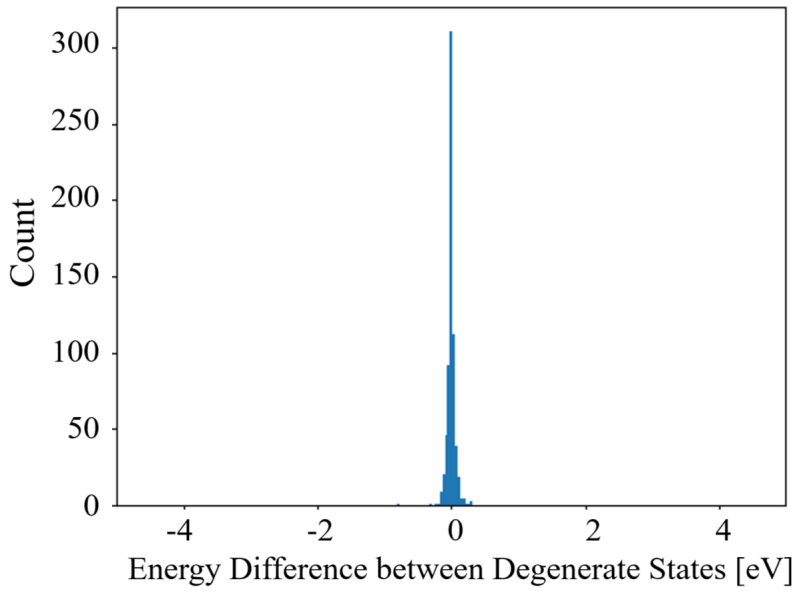

Figure R11 The histogram of energy difference between 663 pair of degenerate states. The mean energy difference is 0.03 eV with a high $R^2 = 0.99$

We have added this information to the supplemental material (Supplementary Fig. S4, Fig. S14 and Fig. S12).

A deeper statistical analysis of the results beyond the mean absolute error (MAE) is necessary. The presentation of a single band structure is not sufficiently representative of the model's performance. It is crucial to analyze how the GW correction varies across different k-points and bands and whether the model accurately respects symmetries at high-symmetry points. What is the MAPE? What is the maximum error and for which kind of materials?

We have addressed to how the GW correction varies across different k-point and bands and whether degeneracy is respected in the previous answer.

We respectfully disagree with the referee's suggestion to use the mean absolute percentage error (MAPE) as a main metric for our model's performance. While MAPE is an easy and straightforward prediction metric for some forecasting models, it is not suitable for evaluating our model in this context. Unlike previous machine learning predictions of GW correction or band gap [2,9,10] that focused only on insulators, our GW training and test sets are much more general, including both insulators and metals. Consequently, a substantial portion of the GW corrections in our training set are close to zero, making MAPE a misleading metric that becomes numerically unstable as the correction approaches zero. As illustrated in the Fig. R12, we examined how MAPE changes with respect to the absolute value of GW corrections above specific energy thresholds. For instance, although our downstream regression achieves a low MAE of 0.1 eV for GW corrections in the test set ranging from -5 to 5 eV, MAPE becomes unreliable for corrections near 0 eV. This results in an unreasonable overall MAPE of 54% for the overall test set. However, if we exclude absolute GW corrections below 0.8 eV, the MAPE quickly drops below 5%. Given

this situation, using an unstable metric like MAPE is unreasonable. Alternatively, as the solid line illustrate in the Fig. R12, the R^2 metric offers a more stable and appropriate evaluation of our model's performance since R^2 is renormalized by the variance of the data rather than the absolute data values, making it more stable to evaluate the model across different energy ranges, as shown in the Fig. R12. Therefore, we use R^2, which has also been widely employed in the past GW correction studies[2], to monitor the training process and evaluate our model's performance.

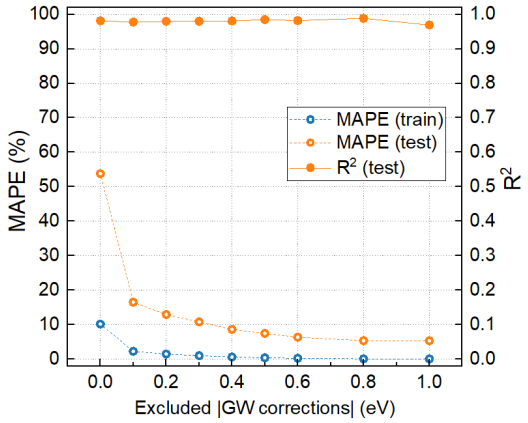

Figure R12 The evolution of different metrics with respect to different energy range of prediction. The orange (blue) circles represent the mean absolute percentage error on the test (training) set. The orange dots represent the R^2 on the test set. The x-axis represents the dataset excluding GW corrections from 0 eV.

More details are needed regarding the training process, such as how the training progresses with increases in the number of data points and epochs. This information is crucial for understanding learning dynamics.

We appreciate the reviewer's suggestion to provide more detailed metrics regarding the training process. Following this advice, 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 Fig. R13 (a) below shows the R^2 performance of the model on the training and validation sets, achieving high R^2 values of 0.93 and 0.91within 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 Fig. R13 (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 Fig. R13 (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. We have added this analysis the supplemental information.

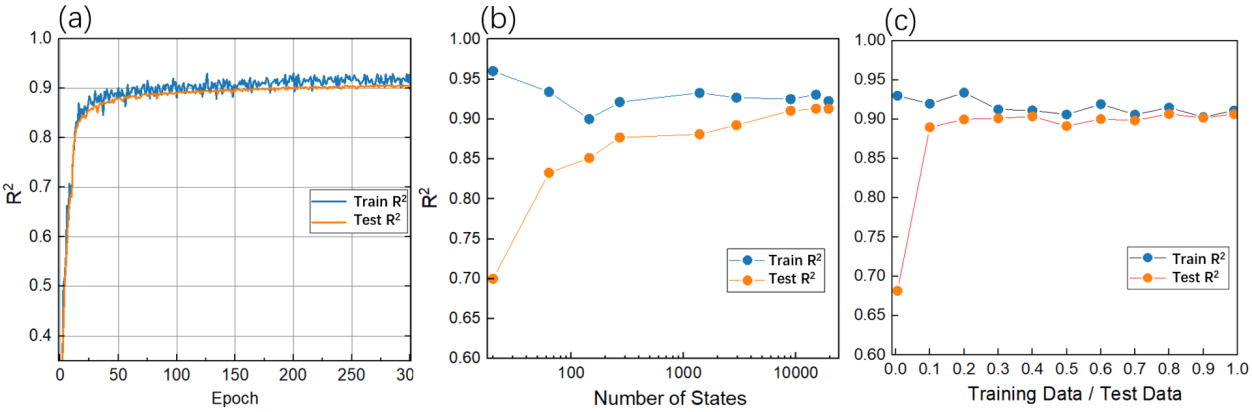

Figure R13 (a) Training dynamics of VAE for KS states with respect to different number of epochs. (b) VAE R^2 performance with respect to different size of dataset. (c) VAE R^2 performance with different ratio of training/test data. The total dataset includes 20,000 KS states.

For the downstream supervised prediction, we also conduct more benchmarks on our model. Fig. R14(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. Fig. R14(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 [11], as shown in Fig. R14(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 R^2 score of our model is 0.96 on the training set and 0.94 on the test set.

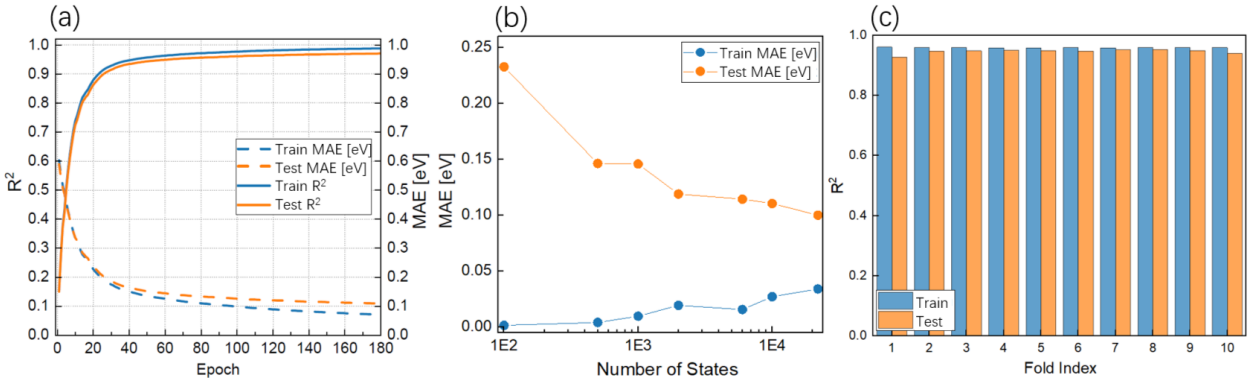

Figure R14 (a) Training dynamics of downstream NN for GW corrections with respect to different number of epochs. The orange (blue) dashed lines represent test (training) MAE. The orange (blue) lines represent test (training) R^2. (b) Downstream NN R^2 of performance for GW prediction with respect to different size of dataset. (c) 10-folds cross-validation of downstream NN for GW prediction.

We also appreciate the reviewer's question regarding the large outliers 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.

We have added this information to the supplemental material (Supplemental Fig. S5-6).

Given these concerns, I am currently unable to recommend this manuscript for publication in Nature Communications as it stands. Enhancements in the model flexibility and a more comprehensive evaluation of its predictive capabilities and training dynamics, are needed before it can meet the publication standards.

We hope that we have adequately addressed the reviewer's concerns and that our manuscript is now suitable for publication in *Nature Communications*.

Reply to referee 3:

In this work, the authors present a useful methodology for learning single-particle Kohn Sham wavefunctions (KS) from density functional theory (DFT). They use an unsupervised VAE encoding scheme to learn a low-dimensional latent space representation for the KS wavefunctions, which can then be used to decode accurate predictions for KS wavefunctions outside of the training data set. They also wrap a GW band structure prediction learning algorithm around their KS learning algorithm, and find good agreement between its predictions and first principles GW band structure predictions. The unsupervised nature of the KS learning algorithm is perhaps the most intriguing aspect, as the authors have mentioned that most other DFT learning algorithms require heavy-handed feature selection. I believe the authors **make significant contributions and present their results in a convincing manner in this manuscript**.

I have two comments to make regarding the manuscript, after which being addressed I believe it would be suitable for publication.

We thank the referee for his or her careful evaluation of our paper and recommendation for publication pending some clarifications and benchmarking. In particular, the referee inquired whether this algorithm provides significant improvements in the study of ground state density properties, which is the physically relevant quantity in DFT. After more benchmarking and testing regarding to ground state density, our model can indeed reproduce the ground state charge density suggesting it is a useful representation for other downstream ML purposes. We address the referee's questions in greater detail below.

The first is that the authors' have focused, in this paper, on reproducing the single-particle KS wavefunctions. This is evident by looking at their cost function, which is a least-squares over single particle wavefunctions. The issue therein is that the KS wavefunctions themselves are not a physically relevant quantity in DFT, the electronic density is. **It is unclear, then, whether the learning algorithm performs well in reproducing the ground state density properly**. Furthermore, since most DFT XC functionals involve complex non-local integrals over the density, it is also unclear whether this algorithm provides any significant improvements in the ground state density properties relevant to DFT. **I do understand that the authors focus in this manuscript is the reproduction of single-particle band structures**, for which the learning algorithm may be suitable, **but it seems warranted to discuss errors in densities and total ground state DFT energies relative to the first principles data.** This should be easy to manage, since the authors must have already computed the first principles DFT energies and densities en route to the results they have in their manuscript.

We appreciate the reviewer's thoughtful question regarding the extension of our model to physically relevant observables in DFT, particularly the accurate reproduction of the ground state charge density by the VAE. We address this in two ways: firstly, we look at how the representations of the KS wavefunctions (VAE-KS) reproduce the charge density, and secondly, we look directly at VAE representations of the charge density (VAE-charge).

Firstly, we investigate if the KS states generated by the VAE (VAE-KS) can reproduce the charge density. We pass all occupied states of a material to VAE-KS and obtain the reconstructed KS state $\varphi'_{nk}(\boldsymbol{r})$, then sum over occupied states to form the reconstructed charge density:

$$\rho'(\boldsymbol{r}) = \sum_{nk}^{occ} |\varphi'_{nk}(\boldsymbol{r})|$$

The R^2 of the VAE-reconstructed wavefunction is 0.92 across the test set. Remarkably, the VAE-reconstructed wavefunctions reproduce the ground state charge density with an even higher $R^2 = 0.99$ even though the total charge density is never used in the loss function, suggesting that the VAE representations of the KS states can effectively be used for prediction of ground state properties associated with the charge density. This improvement may arise from cancellation of error when summing across multiple KS states. We have added a sentence stating this in the manuscript.

Secondly, as noted by the reviewer, for ground state properties, the KS wavefunctions are less physically relevant than the ground state charge density, so it would be helpful to directly evaluate the VAE representation of the ground state charge density. For clarity, we will refer to this as VAE-charge. We have performed additional extended study on the predictive/generative power of the latent space of VAE-charge for ground state observables and material classification.

For independently training 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. The Fig. R15(a) shows the R^2 performance of the model on the training and test 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 Fig. R15(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 being overfitting. Fig. R15(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 general representation mentioned in the manuscript.

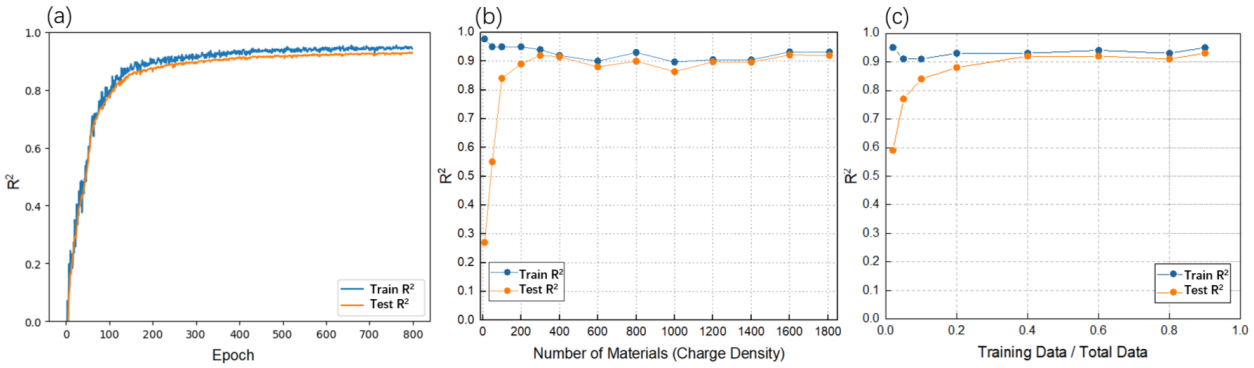

Figure R15 (a) Training dynamics of VAE-charge with respect to different number of epochs. (b) VAE-charge R^2 performance with respect to different size of dataset. (c) VAE R^2 performance with different ratio of training/test data. Total dataset includes the charge density of 2670 2D materials.

Secondly, 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. As the referee suggests, the prediction of the total ground state energy is a good example for the 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 80% for the training set and 20% for the validation/test set. As shown in Fig. R16 (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 Fig. R16 (b).

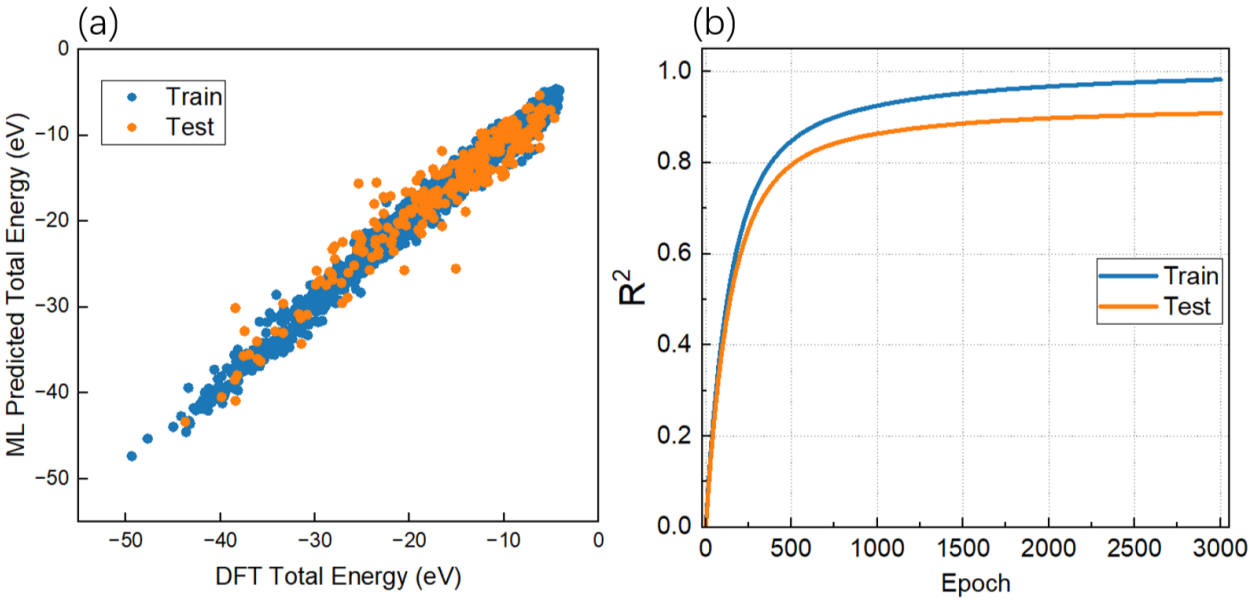

Figure R16. Parity plot comparing the exact calculated values (x-axis) to the ML predicted values (y-axis) of the DFT total energies for individual material. Blue (orange) dots represent training (test) sets. The MAE for the training set and test set are 0.7 and 1.5 eV respectively. (b) Training dynamics of downstream NN for DFT total energies with respect to different number of epochs. The orange (blue) lines represent test (training) R^2.

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 Fig. R17, the ROC AUC score achieves a high value of 0.89, which is comparable with previous ML model exclusively trained for metal/insulator classification [12].

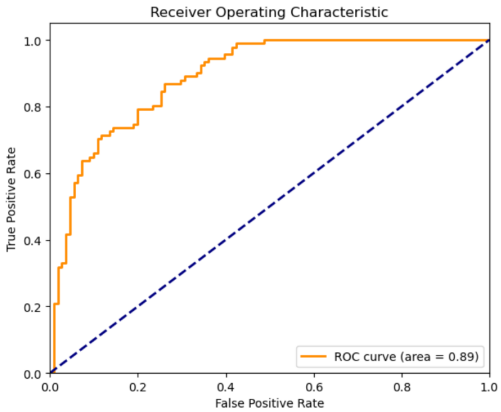

Figure. R17 Receiver Operating Characteristic (ROC) curve for metal vs. insulator classification using VAE charge density representation: The model achieves a high area under the curve (AUC) value of 0.89, indicating strong performance in distinguishing between metal and insulator states.

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.

We have added this information to the supplemental material (Fig. S7, Fig. S13 and Fig. S15)

Secondly, I think some discussion about potential applications to other post-HF corrections may be useful (maybe just a paragraph in the conclusion). For example, using the ML generated wavefunctions in quantum Monte Carlo (QMC). In particular, there is a methodology for developing low-energy Hamiltonians which relies on efficient representations of low-energy wavefunctions called Density Matrix Downfolding [https://arxiv.org/abs/1712.00477]. This method usually involves a lot of user supervision, especially in sampling and constructing low-energy wave functions, but it seems that the method shown by the authors in this manuscript may be able to provide a clean unsupervised way to do this, at least at the DFT level. Given that most QMC wave functions are built from DFT wavefunctions, the connection to the prior should be straightforwards. In general, some discussion on post-HF applications of the VAE wavefunctions would be useful, since many post-HF (QMC, CCSD, CASCI) involve DFT wave functions as a starting point.

We appreciate the reviewer's suggestion to further discuss potential applications to other post-HF corrections. We have added a related discussion on post-HF applications of the VAE wavefunctions, particularly focusing on Density Matrix Downfolding, CCSD, CASCI.

Lastly, I recommend the authors rephrase statements such as "wiggles in band structures" to something more appropriate and precise. Similar kinds of statements are made throughout the manuscript.

We have modified it in new revision.

Code runs properly.

Thanks!

[1] A. Zadoks, A. Marrazzo, and N. Marzari, arXiv preprint arXiv:2403.01514 (2024).

[2] N. R. Knøsgaard and K. S. Thygesen, Nature Communications **13**, 468 (2022).

[3] A. Baratz, G. Cohen, and S. Refaely-Abramson, arXiv preprint arXiv:2404.11980 (2024).

[4] J. Zang, M. Medvidović, D. Kiese, D. Di Sante, A. M. Sengupta, and A. J. Millis, arXiv preprint arXiv:2403.15372 (2024).

[5] Y. Luo, D. Desai, B. K. Chang, J. Park, and M. Bernardi, Physical Review X **14**, 021023 (2024).

[6] M. W. Diederik P Kingma, arXiv (2013).

[7] M. Lin, Q. Chen, and S. Yan, arXiv preprint arXiv:1312.4400 (2013).

[8] S. Schubert, P. Neubert, J. Pöschmann, and P. Protzel, in *2019 IEEE intelligent vehicles symposium (IV)* (IEEE, 2019), pp. 653.

[9] A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, and A. K. Singh, Chemistry of Materials **30**, 4031 (2018).
[10] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, Physical Review B **93**, 115104 (2016).
[11] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009), Vol. 2.
[12] A. B. Georgescu, P. Ren, A. R. Toland, S. Zhang, K. D. Miller, D. W. Apley, E. A. Olivetti, N. Wagner, and J. M. Rondinelli, Chemistry of Materials **33**, 5591 (2021).