Reviewer 1

Q1: The authors assume that there are physics bases $\boldsymbol{z}$ that yield a linear relationship $\boldsymbol{y} = A\boldsymbol{z}$. In addition, they assume the unobservable base is also a linear transformation of the observable one, i.e., $\boldsymbol{z}_U = B\boldsymbol{z}_O + \boldsymbol{e}$, where $\boldsymbol{e}$ is relatively small noise. I wonder this assumption is general for arbitrary physical systems or just for fine the examples presented in this paper.

Response: In this paper, we assume physical bases ϕ are known. Then, for an arbitrary system, a linear relationship is sufficient to represent the system $\boldsymbol{y} = A\boldsymbol{z} = A\phi(\boldsymbol{x})$ and the nonlinarity only happens in ϕ. Further, for the equation $\boldsymbol{z}_U = B\boldsymbol{z}_O + \boldsymbol{e}$, we assume there is a linear relationship between the observed and hidden quantities. This may not be generally true for all physical systems and we will elaborate other relationships in the future work.

Q2: Because the proposed architecture is complicated, I wonder all the components are truly necessary. The authors claim the usefulness of each component indirectly, by comparing other baselines (e.g., the usefulness of physics bias with comparing ResNet, that of VNN with comparing SINDy, ...), However, a more systematic ablation study should be included in the paper (or Appendix if the page is not sufficient), to clearly support the necessity of the complicated ATN structure.

Response: We conduct the ablation study. Specifically, we remove the sparsity of PNN, skip-connection of VNN, and mini-max training, respectively. Then, we test the models in different systems and achieve similar results. For example, for 85-node system and compared to the complete ATN model, the mean of $nTVE(\%)$ has a increase of 35%, 22%, and 18%, respectively. This shows that the three components have great impacts on learning the correct physics. Further, sparsity plays the essential role as it restricts the representation of the physics.

Q3: It seems that the following paper is highly relevant: "Y Yin et al. Augmenting physical models with deep networks for complex dynamics forecasting. ICLR 2021." This paper also decomposes the unknown physical system as a known physics part (however whose parameters are unknown) and the neural network part that fills the remainder that cannot be explained by the physics-based one.

Response: The authors in Yin's paper also assume a physical and an augmented neural networks to represent the underlying dynamical equations of the physical model. However, their method has the following disadvantages and is unqualified for our setting. (1) their design of restricting the output of augmented NN is to minimize its norm. For our dataset, this restriction easily causes overfitting of the physical NN. Specifically, the physical parameters will be mis-estimated to enable the physical NN to fit the part that augmented NN can't approximate. Correspondingly, in our model, we also restrict the VNN's representation using a skip connection. However, we add a mini-max learning scheme to maximize the dissimilarity between the outputs of PNN and VNN. Thus, PNN can't represent the noise that should be handled by VNN, leading to a better output boundary between PNN and VNN. (2) they consider noiseless dataset while the noise can further deteriorate the correct boundary. (3) they only focus on low-dimensional data but don't present results for high-dimensional data.

Q4: It seems that the proposed ATN already knows the appropriate physics bases $z = \phi(x)$, referring to the explanation on page 3 ("we fix the parameters from x_O to z_O") and equations (1 – 4) (no x and ϕ in the objectives). However, on page 7, the authors state that "ATN and SINDy can both learn physical bases", which seems that there is a searching procedure of physics bases from a set of predefined operations. Can the authors elaborate it more clearly?

Response: In this paper, the physical bases are assumed to be known and we need to estimate the parameters. On page 7, our claim should be "ATN and SINDy have the ability to estimate parameters of physical equations."

Q5: There are some inconsistent or confusing notations in the paper, e.g., on page 3 "Correspondingly, we denote $z = [z_1, z_2]^\top$, where z_O are physical bases...", on page 4, $y_O = g_p(z_O) + g_v(g_p(z_O))$ should be y rather than y_O (because g_v is included in the expression, and consistency with (1 – 4))? In addition, Figure 1 is not clear because all subscripts of variables are 1, 2, ... rather than O or U.

Response: We re-clarify our notations. On page 3, we denote $\boldsymbol{z} = [\boldsymbol{z}_\mathcal{O}, \boldsymbol{z}_\mathcal{U}]$. On page 4, the notation should $\boldsymbol{y}_\mathcal{O}$ rather than $\boldsymbol{y}$. The reason is that $\boldsymbol{x}_\mathcal{O}$ and $\boldsymbol{y}_\mathcal{O}$ are the observed input and output measurements of the physical system and we don't have the unobserved measurements $\boldsymbol{y}_\mathcal{U}$ for training. Further, the unobserved measurements $g_v(g_p(\boldsymbol{z}_\mathcal{O}))$ should still contribute to the observed measurements $\boldsymbol{y}_\mathcal{O}$ due to the possible physical connections between unobserved nodes and observed nodes. Finally, we will modify Fig. 1 for better clarification.

Reviewer 2

Q1: Beginning of section 3.1 The author frame the problem as a graph with node set V and edges $V \times V$. Each node in V contains (x, y) with the relationship $y = f(x)$. Does this mean that (x, y) are associated with each node and for

different nodes there are different (x, y)?

Response: $\boldsymbol{x}$ and $\boldsymbol{y}$ represent the system input and output for system nodes and each elements in $\boldsymbol{x}$ and $\boldsymbol{y}$ represent one nodes' measurements.

Q2: Then the assumption is made for some nodes being observed and some not observed. $V = O \cup U$. The the new symbols are introduced $[y_O, y_U] = f([x_O, x_U])$. Seems to me that y_O is a set instead of variables representing one specific node. Otherwise this equation $[y_O, y_U] = f([x_O, x_U])$ does not make sense. If so, then the (x, y) in my point Q1 are variables, not set and here y_O is a set of variables. Please clarify this point.

Response: Yes, your point in Q1 is correct. Please see my answers in Q1.

Q3: There could be a typo here: "Correspondingly, we denote $z = [z_1, z_2]^T$, where z_O .." What are z_1 and z_2?

Response: Yes, it's a typo. We denote $\boldsymbol{z} = [\boldsymbol{z}_{\mathcal{O}}, \boldsymbol{z}_{\mathcal{U}}]$.

Q4: ϕ the last paragraph of section 3.1 represent the basis coefficients for some known basis functions? Are the values of ϕ learned end-to-end in this network? By learning PNN?

Response: In this paper, we assume the physical bases ϕ are known.

Q5: Eq(3) the variable Y. $Y = 0$ if sample pair is similar and $Y = 1$ otherwise. In Eq(3) we only pass in one z_O, not clear where the 'pair' comes from. Perhaps it is due to the confusion I brought up in Q2. If z_O is a set of variables, then how to apply a function to it? Element wise application?

Response: As shown in Fig. 1, $g_{e_1}(g_p(\boldsymbol{z}_{\mathcal{O}}))$ and $g_{e_2}(g_v(\boldsymbol{z}_{\mathcal{O}}))$ represent the pair. $\boldsymbol{z}_{\mathcal{O}}$ is a set of variables and the function represents a neural network.

Reviewer 3

Q1: The beginning of the problem formulation in section 3.1 is misleading. Now it says:"A physical system can be modeled as a directed weighted graph...". This sounds a bit weird because many physical systems are not usually modeled as a directed graph. Instead, it would be more appropriate to say that you handle specific types of physical systems that are modeled as directed weighted graphs.

Response: Yes, we follow your suggestion and change the statement to "In this paper, we handle physical systems that can be modeled as a directed weighted graph."

Q2: Seemingly, at the layer of the problem setting, the main difference from [Li & Weng, 2021] is the presence of a possibly nonlinear function f between x and y. Meanwhile, such a difference is actually not valid because now the prior knowledge of the nonlinear functions used in f is assumed as ϕ. Consequently, in my understanding, there is no effective difference between the problem setting of the current paper and that of [Li & Weng, 2021]. I would like to hear if this understanding is correct or not. Please note that I understand the methods in both papers are more or less different; here I question the difference of the problem setting.

Response: Yes, the problem setting is the same. However, our paper doesn't need to be restricted to the specific physical systems with symmetric Laplacian matrix as the system parameter matrix. Further, for PCNN, the training may be reinitialized multiple times to obtain a good result. The reason is pointed out in our paper: the direct training can't guarantee the correct boundary between the physical output (e.g., the output of PNN) and the virtual output (e.g., the output of VNN). Thus, in this paper, we propose to utilize mini-max game to solve this problem.

Q3: What are "the remaining quantities" that appear at the second line of section 3.3? Do they refer to x's (or z's) on unobserved nodes?

Response: Yes, your are right. The remaining quantities refer to the hidden measurements.

Q4: I fully agree with the need for restricting the output of VNN and encouraging PNN to output as much as possible. In fact, similar ideas have been already investigated in the context of physical system learning in literature such as [Yin et al., 2021] and [Takeishi & Kalousis, 2021].

Response: The authors in Yin's paper also assume a physical and an augmented neural networks to represent the underlying dynamical equations of the physical model. However, their method has the following disadvantages and is unqualified for our setting. (1) their design of restricting the output of augmented NN is to minimize its norm. For our dataset, this restriction easily causes overfitting of the physical NN. Specifically, the physical parameters will be mis-estimated to enable the physical NN to fit the part that augmented NN can't approximate. Correspondingly, in our model, we also restrict the VNN's representation using a skip connection. However, we add a mini-max learning scheme to maximize the dissimilarity between the outputs of PNN and VNN. Thus, PNN can't represent the noise that should be handled by VNN, leading to a better output boundary between PNN and VNN. (2) they consider noiseless dataset while

the noise can further deteriorate the correct boundary. (3) they only focus on low-dimensional data but don't present results for high-dimensional data.

As for Takeishi's paper, the authors also decompose the learning model into the physical and the viartual parts. Futhre, they also try to find the correct output boundary by regularization. Specifically, they regularize the virtual NN by minimizing the discrepancy between the learning model and the "physics-only" reduced model to encourage the learning model to guarantee physics superiority. Secondly, they introduce physics-part recognition network to restrict the use of physics. In our paper, we have similar ideas for restrictions for both PNN and VNN. However, their methods require strict assumptions on the variable distributions and the known physical knowledge. On the other hand, we don't have these assumptions and utilize simple but effective mini-max game for learning.

Q5: In section 3.3, the skip-connection of the two neural nets is "proposed," while it is just a frequent practice.

Response: Yes, we will change "propose" to "employ".

Q6: The regularization in equation (1) to encourage g_p to directly approximate y is reasonable. I note that a very similar idea was adopted in [Yin et al. 2021], where they regularize the norm of the non-physics part (i.e., g_v if the notation is aligned to the current paper). I think the two approaches, i.e., minimizing $||y - g_p(z)||$ and minimizing $||g_v(g_p(z))||$, are almost the same thing eventually after fitting gets enough good. However, in the early phase of training, since the models are not well fitted yet, perhaps there might be some practical difference.

Response: I agree that the two regularization terms are very similar. However, we emphasize here only utilizing the regularization may hardly guarantee the correct estimation of the underlying physical parameters. Specifically, restricting the capacity of VNN may drive the PNN to output the uncertainties and deliver incorrect PNN parameters. Thus, we need to further restrict the PNN. Essentially, this process is to find the correct output boundary between the PNN and VNN, which is never solved before. In our paper, we employ a mini-max game to gradually uncover the boundary accurately. The game mechanism encourage the dissimilar distributions between the outputs of PNN and VNN and thus, leading to correct boundary.

Q7: Does y without subscripts in equation (1) mean $[y_O, y_U]$, similarly to z and x? Maybe I missed something but could not find an exact definition.

Response: We change $\boldsymbol{y}$ to $\boldsymbol{y}_{\mathcal{O}}$ in Eqn. (1).

Q8: The most puzzling thing is that in section 3.4, g_v suddenly starts to take z_O directly as an argument, whereas it used to take $g_p(z_O)$ as an argument in section 3.3. What are the definitions of the domain and the range of the functions around here?

Response: It's a typo. g_v should take in $g_p(\boldsymbol{z}_{\mathcal{O}})$.

Q9: I could not see the point of introducing matrix B in section 3.4. Does it appear somewhere in the model?

Response: B is to point out the possible relationship between the observed and the hidden measurements. We assume its a linear mapping B for explanations. Then, the parameters learned in PNN should be a function of A and B, representing a reduced physical system within only observed nodes.

Q10: What is "large common knowledge", at the fourth line of the fourth paragraph of section 3.4?

Response: The common knowledge refers to the distribution similarity here. We will add the explanations.

Q11: What is the relation between the proposed adversarial loss and the well-known Wasserstein distance based on the Kantorovich-Rubinstein duality (see, e.g., the paper of Wasserstein GAN)? Do you have any observations on the advantages of the proposed loss, compared to the Wasserstein distance?

Response: The Wasserstein distance measures the distance between two distributions while our proposed loss function is an Euclidean-based loss to measure the distance of sample pairs in contrastive learning. In our case, since we compare the distance of the output pairs between PNN and VNN, we employ the Euclidean-based loss. Further, as in the mini-max game, embedding networks are added to convert the output pairs of the NN twin to a embedding space, the proposed sample pair loss is enough to capture the distribution similarity for the output pairs.

Q12: What is the actual use of Y in equation (3)? In my understanding, the idea here is to maximize dissimilarity between $g_p(z)$ and $g_v(z)$ for each z, and thus $Y = 1$ is always used. Then, there is no need to consider the case of $Y = 0$. Is this correct?

Response: our paper utilizes a mini-max training scheme to learn the representation with maximal dissimilarity. We can't do the direct maximization because we can't not directly obtain the correct embedding neural networks to find the similarity of outputs between PNN and VNN. In another

word, we need a minimization process to specially train the embedding neural network and learn the similarity of the output of PNN and VNN, i.e., $Y = 0$ in the minimization process.

Q13: What ϕ exactly was given to the proposed method in each experiment?

Response: In our experiments, we set $\phi = [x^2, \sin x]$ for power systems. We will add the explanation to the paper.

Q14: In the experiment, are the baseline methods also given the knowledge of the base functions, i.e., ϕ? If not, then the comparison is not really fair.

Response: Yes, all the baseline methods include the knowledge of base functions.

Q15: The authors say that one of the baselines, PCNN, "doesn't include a clear design of LASSO regularization." However, in my understanding, [Li & Weng 2021] use lasso for initializing the parameters for fully observed nodes. Do you also use the same lasso-based initialization for PCNN?

Response: We follow exactly the same initialization process as the original paper in PCNN. We admit that PCNN also utilizes LASSO regularization and delete the statement about PCNN in our paper.

Q16: Seemingly, both the reported test results and the hyperparameter selection are based on cross-validation. Are there two different (and maybe nested) cross-validation? This is an important question. I would like to know the difference between the cross-validation for hyperparameter selection and that for generating the final test results.

Response: We utilize the same cross-validation to select hyparameters and generate the results. We fine-tune the parameters based on the validation accuracy and generate the results based on the testing accuracy.

Q17: There are no ablation studies on the effect of the proposed regularizers (in equations (1) and (3)), which makes it difficult to assess the goodness of the proposed method.

Response: We conduct the ablation study. Specifically, we remove the sparsity of PNN, skip-connection of VNN, and mini-max training, respectively. Then, we test the models in different systems and achieve similar results. For example, for 85-node system and compared to the complete ATN model, the mean of $nTVE(\%)$ has a increase of 35%, 22%, and 18%, respectively. This shows that the three components have great impacts on learning the correct physics. Further, sparsity plays the essential role as it restricts the representation of the physics.

Q18: Twin structures of neural networks are mentioned in section 2 and some other occasions, but I could not understand why the proposed method also falls in this category, because nothing seems to be twin in the proposed method (PNN is linear, and VNN is nonlinear).

Response: The twin may not need to be symmetry. We utilize twin here to simply indicate that the two networks are coupled to achieve the same target to minimize the total regression error.

Q19: Below, minor points. Unnecessary "and" slips in at the third-last line of page 1. "Target is the identify ..." (the first line of page 2). $z = [z_1, z_2]^T$? (the lower half of page 3). "lower nose level" (the seventh line of page 8)

Response: we have addressed all the minor errors.