nature portfolio

Peer Review File

Refining Potential Energy Surface through Dynamical Properties via Differentiable Molecular Simulation

Corresponding Author: Dr Kuang Yu

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)
The work presents a very interesting idea of basically fitting NN potentials as force fields to the experimental data, in this case to density and IR spectra. This is pretty much in the current trend of training NNs to make simulations better, e.g., with respect to experiment, rather than simply fitting NNs to get smaller test errors or just represent the reference DFT method.

While interesting, it is not clear how practical it is and some claims are not supported. Also, it needs some technical clarifications:

1. How practical would it be to train such potential? I.e., what is the actual cost of training in wall clock time?

2. Since it is so much effort to train & there are apparently many stability issues, isn't it simpler/better to take some universal/coupled-cluster level NN potential and run MD/RPMD?

3. The claim that their DFT-based ML potential is promoted towards the CCSD(T)/CBS accuracy is an exaggeration. The errors are huge (even for intermolecular interactions the erros are of 0.49 kJ/mol/atom!). Again, why not take the potential trained on those CCSD(T)/CBS data and run IR?

4. Connected to points 2-3, how does the authors' approach compare to the potential trained on CCSD(T)/CBS and used for MD/RPMD?

5. In general, it is amazing that with such short trajectory length of, e.g., 0.3 ps, the network is able to learn spectra needing much longer propagation time and not overfit with so many parameters. Isn't then the results of this network better when used only for many trajectories of 0.3 ps length rather than running very long trajectories? Please also clarify why you think such a short trajectory length works.

6. Technical question, please clarify why the reweighting technique is necessary if the initial conditions are already sampled from NVT MD.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The manuscript reports on the development of end-to-end differentiable MD simulations, which can be used to refine the underlying potential energy surface to reproduce ensemble properties, and therefore to fit experimental targets properties. Unlike the usual approach of fitting machine learning interatomic potentials, which uses microscopic observations, here macroscopic/experimental observations are used.

After introducing the method, which, as usual in this field, is very vague (we will never know what exactly a "simple truncation" or a "proper damping" scheme is)

The value of the work is well demonstrated by the refinement of the water EANN model - I found that reproducing the RDF (although the OH is quite far from being "perfectly reproduced", as the authors claim) and the self-diffusion constant is impressive. It is concerning, however, that the total errors on the CCDS(T)/CBS dataset are unchanged, just distributed differently. One wonders how much has been gained and lost in the process. For example, would ice(s) be still well described by the model? Did any spurious minimum configurations appear? The fact that the authors could run stable MD simulations with their refined model (please confirm if this is indeed the case) indicates that perhaps the refined PES is still fine, but more the reader will need more rigorous proof of this.

Very enlightening would have been to show how much better the q-SPC/Fw model can get if experimental (rather than artificially generated) targets were used in refining it.

The main criticism of the paper is its length. Although there is no formal length limit of a Communication, the manuscript feels a lot longer than the suggested 5,000 words. I found that the sections "validating" the method, i.e. refining the LJ fluid and the q-SPC/Fw water model are essentially a trivial task - especially so that the perturbation of the model parameters is minimal, so the expected problems (multiple local minima, overfitting etc.) are not even expected present.

Finally, if the authors are given a chance to revise their manuscript, it should be very carefully checked for grammar and typos.

My recommendation is major revision with a much more thorough validation of the EANN water model, which should also include the consolidation of the content of the manuscript and the exclusion of the trivial validation part.

(Remarks on code availability)
Due to time limits, I haven't downloaded the code, but I can confirm that the code is uploaded to a repository and it is documented to the degree that it can be verified.

Reviewer #5

(Remarks to the Author)
Reviewer's comments on NCOMMS-24-39841

In the present work titled "Refining Potential Energy Surface through Dynamical Properties via Differentiable Molecular Simulation" the authors have developed and demonstrated an effective approach to infer microscopic molecular interactions from spectroscopic data using dynamic gradient-based optimization techniques, significantly contributing to the accuracy of potential energy surface models. Instead of predicting dynamical properties using bottom-up machine learning potential, the authors focus on the inverse problem of refining the underlying potentials using dynamical data, especially spectroscopic data. Using water cluster as test case, the authors demonstrate that combining thermodynamic and spectroscopic data can significantly enhance the accuracy of DFT-based machine learning potentials. Their approach results in more reliable predictions of various properties as radial distribution function, diffusion coefficient, and dielectric constant.
In my opinion, the work appears to be insightful and well-researched. The methodology and collected results are valuable, significantly enhancing the accuracy of potential energy surface models. However, the test case might be limited, and therefore, the proposed approach could suffer from a lack of generality.
I believe that this work is potentially suitable for publication in the Nature Communications journal, after the Authors have addressed my comments reported below.

Comments:

1) The authors could include a more detailed comparison in terms of accuracy, computational efficiency, scalability, and applicability to different systems with existing approaches to better appreciate the novelty and understand the potential limitations of the method.

2) In this work, the authors use a water cluster as a simple example to demonstrate the methodology. Did the authors consider extending the application of the technique to a wider range of systems to test the robustness and generalizability of the method?

3) Implementing automatic differentiation and running large-scale molecular dynamics simulations can be computationally

intensive. Could the authors discuss the scalability of the proposed methodology for very complex or heterogeneous systems?

4) Considering that the accuracy of the results also depends on the quality and precision of the experimental data used for calibration, and that any errors or uncertainties in the experimental data could have a negative impact on the accuracy of the final model, it is important to understand how the authors select their data. What criteria do they use to choose the experimental data? What considerations guide their selection process? The authors should clarify this point better.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors satisfactorily addressed all my previous comments. It is a nice work and might be useful in some applications and inspire related methods.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The manuscript has undergone considerable improvement and the authors have addressed my concerns. I recommend publication. For the record, I am not sure whether asking for ChatGPT's help is a good or a bad thing when writing a scientific article. Probably not, but time will tell. In any case, I hope the discussion were not written by it and neither were the results generated by it.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
The authors have thoroughly addressed the questions and have also made slight modifications to some paragraphs to incorporate our comments. In my opinion, the article can be published in its current form.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career

Researchers who co-review manuscripts.

(Remarks on code availability)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The work presents a very interesting idea of basically fitting NN potentials as force fields to the experimental data, in this case to density and IR spectra. This is pretty much in the current trend of training NNs to make simulations better, e.g., with respect to experiment, rather than simply fitting NNs to get smaller test errors or just represent the reference DFT method.

While interesting, it is not clear how practical it is and some claims are not supported. Also, it needs some technical clarifications:

1. How practical would it be to train such potential? I.e., what is the actual cost of training in wall clock time?

The total time cost in top-down fitting can be computed as (simulation time + gradient time)* iterations. Here, we are fitting both density and IR spectrum using a single NVIDIA V100 card. For density, the gradient time is negligible compared to the simulation sampling time, which is 2 hours for 100 ps using i-PI if doing resampling. But resampling only happens in about 1/5 of the iteration steps. For IR spectrum, the simulation sampling time is quite short, and the simulation + gradient time is about 0.5 hour for each iteration. *In total, for bulk water, it takes about 90 hours to perform 100 iteration of optimization, which is quite affordable.* (See the newly added 'Methods' part, Page 11 in the main manuscript). In general, the training cost is only 1~2 order of magnitude higher than the forward property calculations, and is not directly related to the chemical complexity of the system.

2. Since it is so much effort to train & there are apparently many stability issues, isn't it simpler/better to take some universal/coupled-cluster level NN potential and run MD/RPMD?

Actually, the training cost of a coupled-cluster level NN potential can be way more expensive compared to top-down refinement. Due to its poor scaling, CC calculations are only affordable for tiny molecular clusters (typically <15~20 heavy atoms). And that is usually not large enough to train a generic bulk molecular potential without a delicately designed range-separation scheme (see our previous work in water and PEG: see Ref 1 and 2). Even in the situations that we can afford the calculation, for any reasonable cluster sizes (say 10~20 heavy atoms), DFT method is orders of magnitude faster than CCSD(T)/CBS. And the time cost for performing ~100,000 CCSD(T)/CBS calculations is certainly more than the 90 hours we spent on top-down refinement. Therefore, the strategy of direct fitting to CCSD(T)/CBS data is probably cheaper only when the training clusters are tiny (say <5 heavy atoms), which is sometimes true for simple small molecules like water, but certainly not the case in a more general setting.

3. The claim that their DFT-based ML potential is promoted towards the CCSD(T)/CBS accuracy is an exaggeration. The errors are huge (even for intermolecular interactions the errors are of 0.49 kJ/mol/atom!). Again, why not take the potential trained on those CCSD(T)/CBS data and run IR?

We apologize for the overclaim and will modify the corresponding statement to make it more moderate. (See page 10, we revise to "DFT-based ML potential can be promoted to higher accuracy comparing to the CCSD(T) reference") Meanwhile, we would also like to emphasize again that directly fitting to CCSD(T)/CBS and run quantum MD is an extremely expensive process (and most of the time impractical) in general. So far, condensed phase PES with such quality (i.e., CCSD(T)/CBS) in production level is only seen in simple systems like water. While the top-down fine-tuning strategy is much more applicable in general and much cheaper to conduct.

4. Connected to points 2-3, how does the authors' approach compare to the potential trained on CCSD(T)/CBS and used for MD/RPMD?

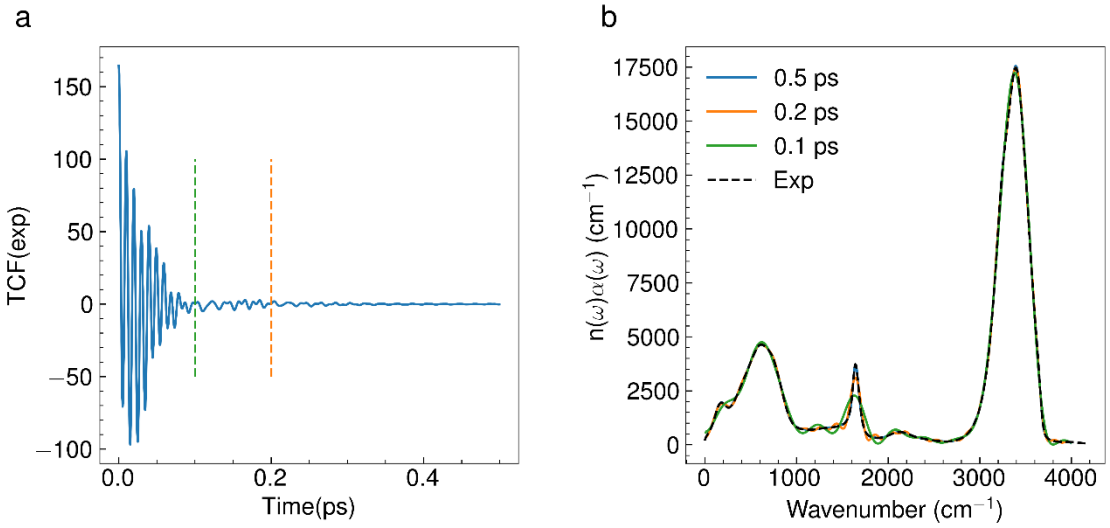
For IR spectrum and density, our model can be better than CCSD(T)/CBS model, because we directly fit these two targets. We can compare with MB-pol (Ref 3), which is a CCSD(T)-level potential energy surface (PES). The IR spectrum using centroid molecular dynamics (Ref 4) and density (Ref 5) of MB-pol is not as good as ours, but for diffusion coefficient and dielectric constant, MB-pol is better (Ref 5), see the table below. However, our method has significant advantages in training cost, as we do not need to generate CCSD(T)/CBS data in massive quantity. Our method thus is also more generally applicable to more complicated systems.

	MB-pol	Our model	Exp
Density (g/cm^3)	1.007	1.00±0.01	0.997
Diffusion coefficient (10^{-9} m^2/s)	2.34	1.99±0.07	2.299
Dielectric constant	68.4	96.1±3.2	78.3

5. In general, it is amazing that with such short trajectory length of, e.g., 0.3 ps, the network is able to learn spectra needing much longer propagation time and not overfit with so many parameters. Isn't then the results of this network better when used only for many trajectories of 0.3 ps length rather than running very long trajectories? Please also clarify why you think such a short trajectory length works.

We apologize for the confusion. First, for the water case, the experimental IR spectrum is smooth thus can be converged relatively quickly. In the panel a of the figure below, we show the inverse Fourier transform of the experimental spectrum (i.e., the TCF), which clearly decays fast. In panel b, we show that 0.2ps TCF already gives a very accurate spectrum. And this is exactly what we do in here: we use many short trajectories (0.4 ps) to fit the IR of water.

Second, we do note that many IR cannot be converged so fast in general. For example, in the case we show in below (see our response to reviewer 5, point 2), the IR signal of the acetonitrile CN group in water is much sharper, thus much harder to converge in the time domain. *In this scenario, we first inverse Fourier transform (FT) the IR to time domain, remove the long-time tail, and FT back, then* ***use the short-time-truncated IR as our fitting target***. Therefore, we always use short trajectories to fit only short-time dynamics, guarantee that it is a fair fitting. (Note that we never use short trajectories to directly fit the full IR!) Once the fitting is finished and we obtain our PES, we always run long time dynamics to compute a fully converged IR and compare it with the experimental results. As shown in below, such strategy works reasonably well. *Essentially, instead of fitting directly to long time dynamics, we are using short time dynamics to refine PES, and use such PES to imply long-time dynamics. Physically, such strategy can work as long as we believe that both short- and long-time dynamics are determined by the same underlying PES, so the improvement of the PES is beneficial for both limits.* We now add more corresponding discussions in our manuscript (page 7 and Supplementary Method 3).



6. Technical question, please clarify why the reweighting technique is necessary if the initial conditions are already sampled from NVT MD.

As the PES parameters (initially set to be θ_0) change, the ensemble distribution changes, so the probability weight of each sample drawn from the θ_0-NVT ensemble should be changed accordingly. Since the target function (i.e., TCF) is in a form of ensemble average, its gradient with respect to θ contains the contribution from the

initial sample reweighting. So the reweighting technique is in principle necessary to evaluate such contribution. Also see the discussions below equation 5 in page 3 for a more detailed explanation.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):

The manuscript reports on the development of end-to-end differentiable MD simulations, which can be used to refine the underlying potential energy surface to reproduce ensemble properties, and therefore to fit experimental targets properties. Unlike the usual approach of fitting machine learning interatomic potentials, which uses microscopic observations, here macroscopic/experimental observations are used.

After introducing the method, which, as usual in this field, is very vague (we will never know what exactly a "simple truncation" or a "proper damping" scheme is)

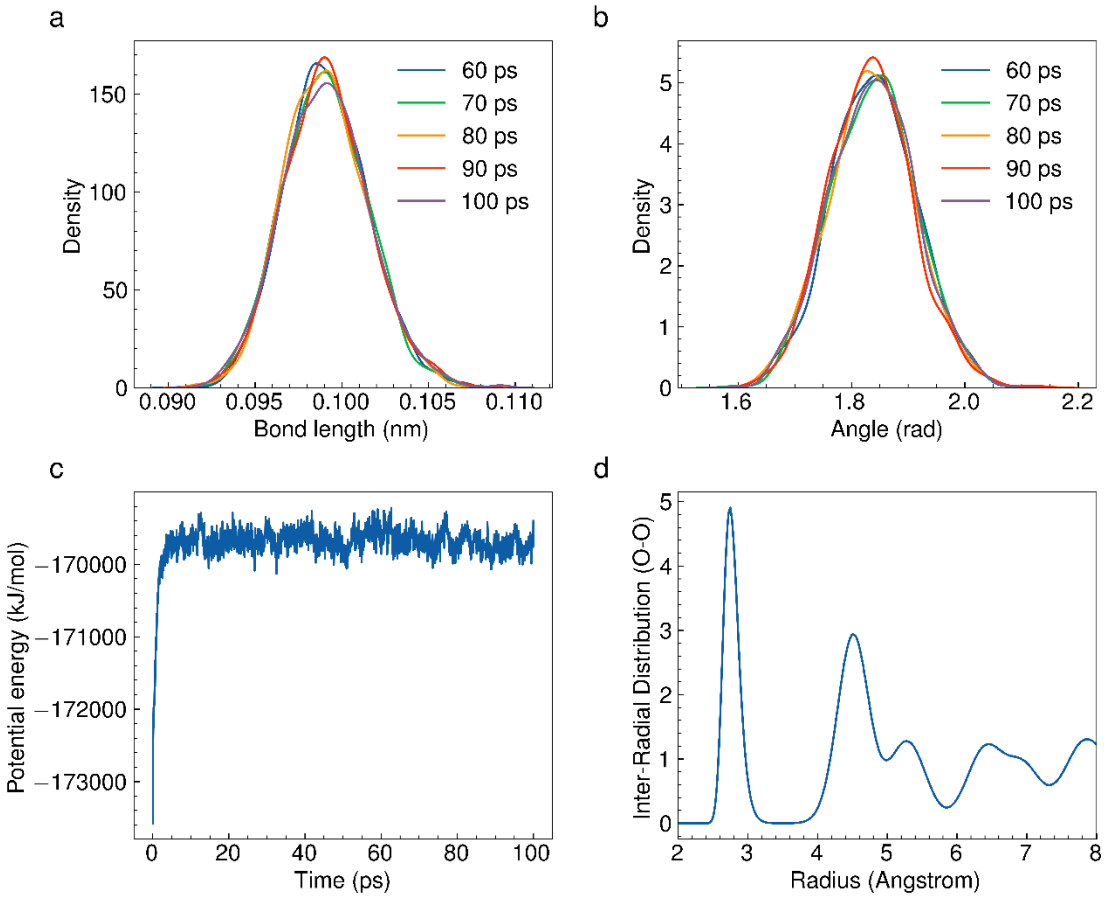
Sorry about that, we revised our manuscript to make it clearer. Here, ‘simple truncation’ means to truncate our TCF at a short time and ‘proper damping’ means to add a Gaussian damping on our TCF curve, which are used to damp the contribution of the long living modes in the object function. The reason to use these schemes is to get stable gradients. As we only analyze the impact of truncation, we will remove the ‘proper damping’ expression in the main text. We also make it clearer in the main text (see page 2 and 4)

The value of the work is well demonstrated by the refinement of the water EANN model - I found that reproducing the RDF (although the OH is quite far from being "perfectly reproduced", as the authors claim) and the self-diffusion constant is impressive. It is concerning, however, that the total errors on the CCDS(T)/CBS dataset are unchanged, just distributed differently. One wonders how much has been gained and lost in the process. For example, would ice(s) be still well described by the model? Did any spurious minimum configurations appear? The fact that the authors could run stable MD simulations with their refined model (please confirm if this is indeed the case) indicates that perhaps the refined PES is still fine, but more the reader will need more rigorous proof of this.

We revised our manuscript to avoid the statement "perfectly reproduced" for OH at

page 8. (The error in the OH RDF is expected though, as the impact of NQE). Conceptually, we can view the refined PES as an approximation to the centroid potential of mean force used in centroid MD, which should work in the calculation of most properties. This PMF will be consistent with energy in the classical limit (for intermolecular interactions) but deviates when there is a strong NQE (for intramolecular interactions).

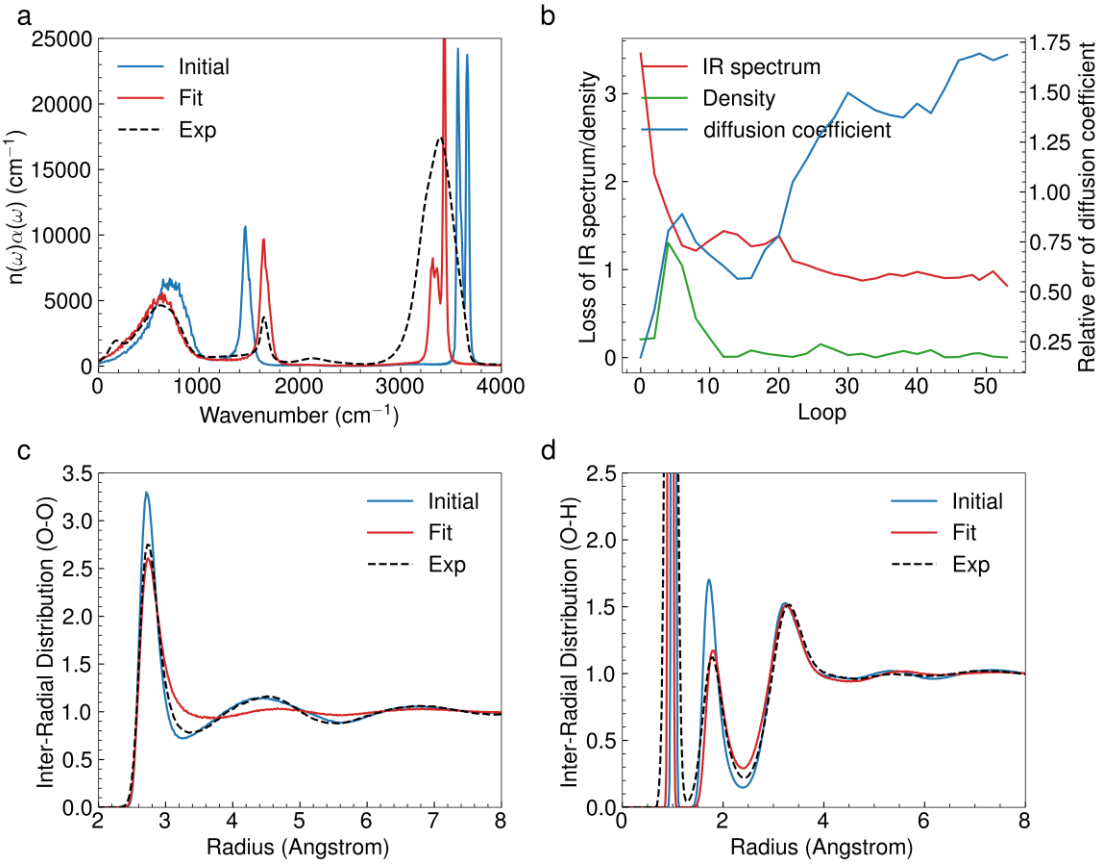
To further verify the quality of the refined potential, we test it on ice Ih. We run a 100 ps NVT simulation on a density of 0.92 g/cm^3 at 241.65 K, it's a stable MD simulation, as evidenced by the stable bond length (panel a) and angle distributions (panel b) at different time points, potential energy at equilibrium (panel c), and OO RDF at equilibrium (panel d). We do not see any spurious minimum configurations in all simulations. This case shows that our refined PES is numerically stable, although the averaged bond length and angle are slightly different compared to experimental result. This is again due to NQE. We do not expect ice can be described better, but it is not ruined. (See page 9)



Very enlightening would have been to show how much better the q-SPC/Fw model can get if experimental (rather than artificially generated) targets were used in refining it.

Following the reviewer's suggestion, we performed such optimization (fitting both IR spectrum and density) using q-SPC/Fw model. The parameters we refined in the q-

SPC/Fw model includes the oxygen charge q, the oxygen σ and ε, the OH bond force constant k_b and length, and the HOH angle force constant k_a and angle. We use point charge to construct the dipole surface, so the O charge is in principle related to the dipole surface. Therefore, we optimize not only the PES, but also the dipole surface. The original and fitted parameters are shown in the table below. The fitted density is extremely accurate (0.9992 versus 1.0176 g/cm^3). The IR fitting results are shown in the panel a of the figure below: the low wavenumber region is fitted accurately, and the positions of other peaks are also correct. The shape of the high-frequency peak is wrong, since q-SPC/Fw uses harmonic bonding terms, which are inherently insufficient. Moreover, as the panel b shows, the error of self-diffusion coefficient increase in the optimization, we also test other properties such as O-O and O-H RDF (panel c and d) and dielectric constant (42 versus 97) using the refined q-SPC/Fw model, the O-O RDF and the dielectric constant also get worse. It seems that there exists a huge gap between the classical force field and the real PES, while we fit one property well, another property can be worse. Such predicament disappears in the case of ML PES, which has strong enough representation ability. (Add this in page 7 and Supplementary Method 3)



	initial	fit
σ (nm)	0.3165492	0.3103229
ε (kJ/mol)	0.6502990368	0.6456576536
k_b (kJ/mol/nm^2)	443153.3808	385342.1689
length (nm)	0.1	0.0941
k_a (kJ/mol/rad^2)	317.5656	371.9439

angle (rad)	1.9547687622336491	2.0709569372027614
q (e)	-0.84	-0.8160

The main criticism of the paper is its length. Although there is no formal length limit of a Communication, the manuscript feels a lot longer than the suggested 5,000 words. I found that the sections "validating" the method, i.e. refining the LJ fluid and the q-SPC/Fw water model are essentially a trivial task - especially so that the perturbation of the model parameters is minimal, so the expected problems (multiple local minima, overfitting etc.) are not even expected present.

Thanks for your suggestion. However, these two cases both elucidate some important techniques that we use in the case of water MLP. The LJ fluid case is mainly used to dissect the performance of thermodynamic and dynamic gradients, as well as to analyze the gradient explosion behavior in details. Meanwhile, the q-SPC/Fw showcase how we use the truncated short-time dynamics as our fitting target to avoid the ill-conditioned long-time differentiation. Also, we now make the q-SPC/Fw case less trivial by conducting a fairly complicated 7-dimensional optimization (see our response in the last point). Therefore, we tend to keep these two cases in the manuscript. However, following the reviewer's suggestion, we do shorten the manuscript by removing many technical descriptions to the supplementary materials, and keep the length of the article roughly below 5000 words.

Finally, if the authors are given a chance to revise their manuscript, it should be very carefully checked for grammar and typos.

We thank the reviewer for raising this point. We now have modified the manuscript using ChatGPT to correct grammatic mistakes.

My recommendation is major revision with a much more thorough validation of the EANN water model, which should also include the consolidation of the content of the manuscript and the exclusion of the trivial validation part.

Reviewer #3 (Remarks on code availability):

Due to time limits, I haven't downloaded the code, but I can confirm that the code is uploaded to a repository and it is documented to the degree that it can be verified.

Reviewer #5 (Remarks to the Author):

Reviewer's comments on NCOMMS-24-39841

In the present work titled "Refining Potential Energy Surface through Dynamical Properties via Differentiable Molecular Simulation" the authors have developed and demonstrated an effective approach to infer microscopic molecular interactions from spectroscopic data using dynamic gradient-based optimization techniques, significantly contributing to the accuracy of potential energy surface models. Instead of predicting dynamical properties using bottom-up machine learning potential, the authors focus on the inverse problem of refining the underlying potentials using dynamical data, especially spectroscopic data. Using water cluster as test case, the authors demonstrate that combining thermodynamic and spectroscopic data can significantly enhance the accuracy of DFT-based machine learning potentials. Their approach results in more reliable predictions of various properties as radial distribution function, diffusion coefficient, and dielectric constant.

In my opinion, the work appears to be insightful and well-researched. The methodology and collected results are valuable, significantly enhancing the accuracy of potential energy surface models. However, the test case might be limited, and therefore, the proposed approach could suffer from a lack of generality.

I believe that this work is potentially suitable for publication in the Nature Communications journal, after the Authors have addressed my comments reported below.

Comments:

1) The authors could include a more detailed comparison in terms of accuracy, computational efficiency, scalability, and applicability to different systems with existing approaches to better appreciate the novelty and understand the potential limitations of the method.

In this work, we used dynamical properties such as transport coefficients and vibrational spectrum to refine the potential energy surface. Such refinement has been extremely difficult using other methods. Other possible optimization strategies include:

1. Gradient-free methods: in such method, we typically sample the parameter space around the current parameter value, and establish a parameter-property relation using a surrogate ML model, then use that surrogate model to guide the optimization (e.g., see Ref 6 and 7). However, it takes at least hundreds of properties (i.e., IR spectrum) evaluations to train the surrogate model every optimization step, which is very inefficient in high-dimensional space.
2. Numerical gradient method: using finite difference method to evaluate the parameter gradient. This is also obviously inefficient: the number of property evaluations in each step scales as $O(n)$, with n being the dimension of the parameter space.
3. Analytical gradient: this approach is essentially impractical for complex computational flow such as IR spectrum, as no closed-form analytical gradient is available.

4. Transforming the gradient into a form of ensemble average and evaluate it in a single MD simulation. This approach is viable for ensemble average and free energy properties (see Ref 8), but has been so far not applicable to dynamical properties.
5. Automatic differentiation techniques: this is by far the most efficient and promising approach. Using the adjoint method, we can evaluate the gradient of the property by effectively running the MD simulations (i.e., the property evaluations) twice. And the cost is irrelevant to the number of parameters, thus is particularly suitable for the refinement of ML potential, which typically contains numerous parameters. Therefore, we believe this is by far the only viable approach for dynamical-property-based ML potential refinement.

We now include this discussion in our manuscript (Page 10).

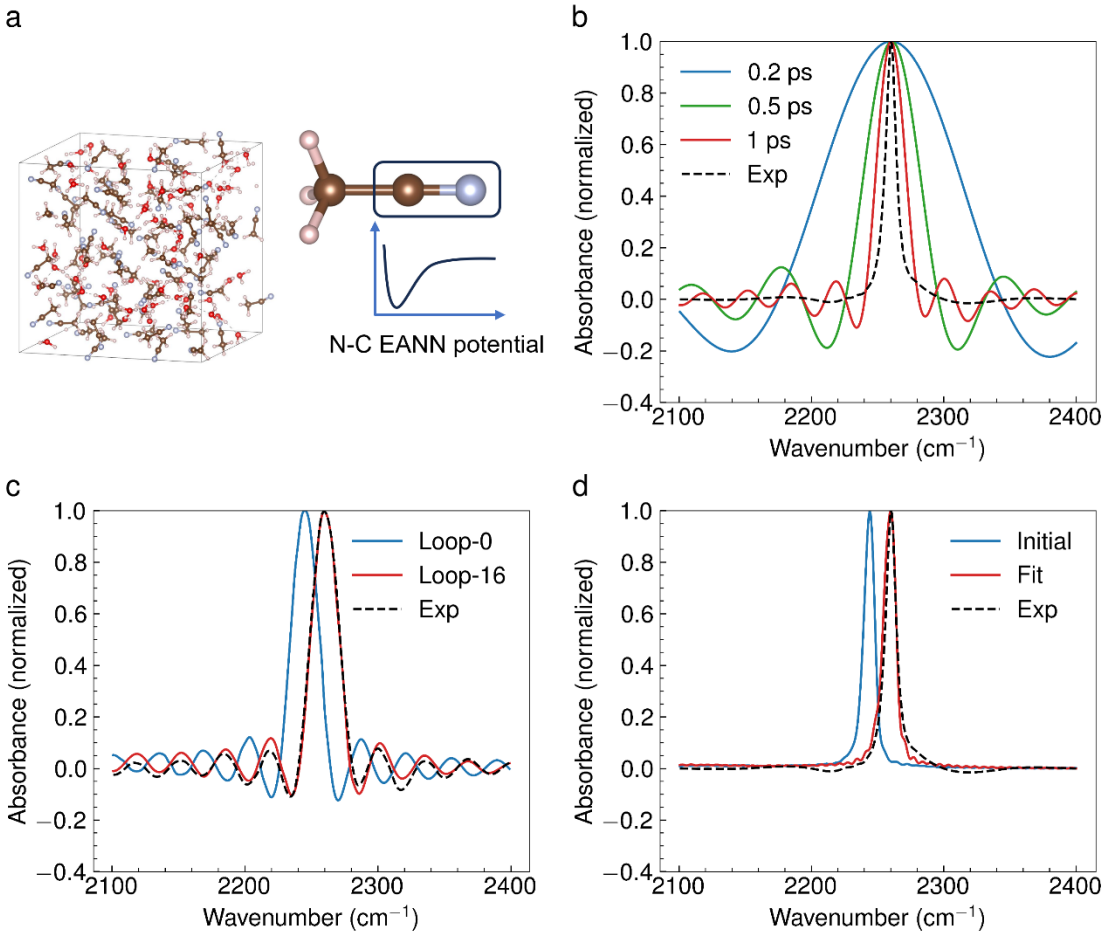
2) In this work, the authors use a water cluster as a simple example to demonstrate the methodology. Did the authors consider extending the application of the technique to a wider range of systems to test the robustness and generalizability of the method?

The main difficulty for the application of this method in complex systems (e.g., complex liquids like battery electrolytes or proteins) is the correct designation of the peaks in the spectrum. If the IR is so complicated and very different from our initial guess, it is very hard to guarantee that the optimization of IR alone will lead to the correct minimum. Physically, it is possible that we match a peak using a phenol OH vibration mode while it is actually an alkanol signal. In future, a multimodal training strategy is more promising, while both bottom-up and various types of top-down targets are combined, so false local minimum can be avoided. While we do believe that dynamical properties such as vibrational spectrum will be an important part of the fitting target.

With that being said, we note that when the peak designation is clear, our method is general and applicable to any systems as long as the spectrum can be computed using normal MD. As a demonstration, we test our method on a slightly more heterogeneous system. Here we fit the C-N vibrational peak in acetonitrile-water mixture. We use EANN model to describe the C-N chromophore group and classical force field for other interactions. Then we aim to refine the EANN potential using experimental IR. The initial guess was obtained by fitting to B3LYP calculations, and the fitting results are shown in the figure below.

Since the CN vibration is only weakly coupled from other modes in water, its lifetime is long (i.e., TCF decays slow), as shown in panel b, the time to converge that peak needs more than 1 ps in TCF domain. Therefore, we use the short-time truncation scheme to avoid gradient explosion. In panel c, we show the fitting quality of the truncated IR spectrum, and in panel d, we show the fully converged IR. As it is shown, in such a heterogeneous case, our method still successfully optimizes the performance

of the PES in IR calculations, proving the applicability of our method in a more general setting. (Add this in page 7)



3) Implementing automatic differentiation and running large-scale molecular dynamics simulations can be computationally intensive. Could the authors discuss the scalability of the proposed methodology for very complex or heterogeneous systems?

In terms of computational cost, our method only needs to backpropagate the dynamics one more time to find the gradient, without any extra memory cost. *Therefore, as long as one can afford the forward calculation (i.e., run a normal MD IR calculation), one can afford the corresponding gradient calculation.* It is noted that it usually takes 10~50 gradient evaluations to converge the parameter optimization, *so a full PES refinement is about one or two order of magnitude more expensive than a single forward IR calculation.* Nevertheless, it is still not prohibitively expensive for any reasonably complicated systems as long as their spectrum evaluation is manageable by MD. (See 'Methods' part, Page 11)

4) Considering that the accuracy of the results also depends on the quality and

precision of the experimental data used for calibration, and that any errors or uncertainties in the experimental data could have a negative impact on the accuracy of the final model, it is important to understand how the authors select their data. What criteria do they use to choose the experimental data? What considerations guide their selection process? The authors should clarify this point better.

The reviewer is correct that we rely on high-quality experimental data, and low-quality data would certainly contaminate the final model. In our work, we choose only the data that is well established by the scientific community as our fitting target to avoid this complication (e.g., the IR spectrum of pure water). Again, in future, such problem can only be solved through multimodal optimization (which we show in small scale as we fit both density and IR simultaneously). The mindset is that while there is error in single type of data, these errors will be counterbalanced by other data of different types. One needs to employ as many types of data as possible to avoid overfitting. The work presented here is a key step towards this future goal. We now add the corresponding discussion in the conclusion section (See page 11).

Reviewer #6 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

References:

1. Yang, Lan, et al. "A transferrable range-separated force field for water: Combining the power of both physically-motivated models and machine learning techniques." *The Journal of Chemical Physics* 157.21 (2022).
2. Chen, Junmin, and Kuang Yu. "PhyNEO: A Neural-Network-Enhanced Physics-Driven Force Field Development Workflow for Bulk Organic Molecule and Polymer Simulations." *Journal of Chemical Theory and Computation* 20.1 (2023): 253-265.
3. Babin, Volodymyr, Claude Leforestier, and Francesco Paesani. "Development of a “first principles” water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient." *Journal of chemical theory and computation* 9.12 (2013): 5395-5403.
4. Palos, Etienne, et al. "Current Status of the MB-pol Data-Driven Many-Body Potential for Predictive Simulations of Water Across Different Phases." *Journal of Chemical Theory and Computation* (2024).
5. Reddy, Sandeep K., et al. "On the accuracy of the MB-pol many-body potential for

water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice." *The Journal of chemical physics* 145.19 (2016)

6. Ge, Pei, Linfeng Zhang, and Huan Lei. "Machine learning assisted coarse-grained molecular dynamics modeling of meso-scale interfacial fluids." *The Journal of Chemical Physics* 158.6 (2023).
7. Befort, Bridgette J., et al. "Machine learning directed optimization of classical molecular modeling force fields." *Journal of Chemical Information and Modeling* 61.9 (2021): 4400-4414.
8. Wang, Lee-Ping, Todd J. Martinez, and Vijay S. Pande. "Building force fields: An automatic, systematic, and reproducible approach." *The journal of physical chemistry letters* 5.11 (2014): 1885-1891.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors satisfactorily addressed all my previous comments. It is a nice work and might be useful in some applications and inspire related methods.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):

The manuscript has undergone considerable improvement and the authors have addressed my concerns. I recommend publication. For the record, I am not sure whether asking for ChatGPT's help is a good or a bad thing when writing a scientific article. Probably not, but time will tell. In any case, I hope the discussion were not written by it and neither were the results generated by it.

Response:
We are using ChatGPT solely to modify the grammar, without altering the semantics of the sentences. After modification, we carefully verified that the original meaning was retained. ChatGPT has not been used to write any content or generate any results in this manuscript.

Reviewer #5 (Remarks to the Author):

The authors have thoroughly addressed the questions and have also made slight modifications to some paragraphs to incorporate our comments. In my opinion, the article can be published in its current form.

Reviewer #6 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.