First, we would like to thank the reviewers for their thoughtful critiques and commendations. We appreciate and are energized by the fact that reviewers found our work to be significant and novel (R2) while positioning itself well with respect to previous work (R2, R3). We have done our best to respond to as many questions and concerns as possible:

While [R3] found the experiments convincing, reviewers also noted that the paper would have been more impactful by including a broader range of experiments that are more realistic / application focussed (R1, R2, R3). We agree the paper would certainly have been strengthened with real-world applications. The intent behind focusing on model problems was to examine the fundamental properties of GHNNs given basic physical systems – leaving applications on complex real-world systems to future work.

[R1] asked about whether metrics were produced using 50 initial conditions from a held-out test set. This was indeed the case and we have made it more clear in our revised paper. Further to this, **[R1] raised some questions about the clarity of the metrics in Appendix D**. We agree we could have done a better job explaining our comparison metrics. For this reason we now have added equations describing the metrics in detail to clear up these concerns.

We want to be clear that we do not wish to claim that **weak regression performs significantly better than state regression (R1).** Rather we claim that "[in our example] weak derivative matching has comparable performance to state regression while requiring substantially less runtime." **[R1] also notes that our claim that "GHNNS perform at least as well as other state-of-the art continuous time models"** is too strong. This is a fair comment and our choice of wording could have been better. We have remedied this claim and have added an extended discussion explaining why we believe GHNNs perform approximately as well as the other modern models for continuous time ODEs while simultaneously learning an underlying energy function.

Thank you **[R1] for sharing an excellent paper from Rudy et. al. [R1] asked about why they appeared to have a lower error on similar problems**. We would like to point out that, while they do a form of state regression in their paper, they use a slightly different metric to compute the performance of their models. More specifically: (i) they compute errors based on the single initial condition they have access to at training time and (ii) they compute a *normalized* error metric that they call the E_F score. We compute a mean E_F score of approximately 0.06 for the weak form regression method and a mean E_F score of 0.27 for the state regression method at a noise of approximately 20% – aligning with their E_F score of approximately 0.23 ± 0.3 for their cubic oscillator (note that we haven't done a full study to validate this E_F score – we only made use of the two training trajectories and the pretrained models we had used to produce Table 1). Thank you **[R1] for suggesting we plot the noisy data on top of a nominal trajectory as was done in this paper**. We agree this would make our exposition more clear and we have added these plots to the revised paper.

[R1] also asked about why Figure 2 in [17] seemed to show lower state errors on a real pendulum than our simulated pendulum. We note that the real pendulum data is significantly less noisy than our simulated data (see Schmidt et. al. "Distilling Free-Form Natural Laws from Experimental Data", Science, 2009). Furthermore, they only model the system for approximately 20 seconds over which time their system can be approximated as energy conserving – making it a good candidate for HNNs.

[R2] points out that theorem 1 is unclear. Thanks for pointing this out, we've updated the proof to make it clear that $g_{ij} = g_{ji}$ in our parameterization. **[R2] also notes that it's not clear if all g_{ij}'s are the same network**. We've updated our paper so that eq. (7) now reads $g_{i,j}(\mathbf{x}_{\setminus ij}) = \mathcal{N}_{i,j}(\mathbf{x}_{\setminus ij})$ to make it clear that these are different networks.

[R4] asked if the method is still applicable in the case m=1. We can set m=1 without changing the method presented here. The number of independent trajectories required will depend on the complexity of the ODE. We noted the number of independent trajectories used for each experiment in the original submission.

[R4] asks if the measurements are assumed to be exact. We assume various amounts of zero mean Gaussian noise on our measurements that are listed in the original submission. We believe our experiments show our methods are effective given noisy measurements.

[R4] also notes that a draw back of weak form regression is that you are restricted to low order quadratures since you cannot evaluate x(t). This is true in the way we have presented our work here. We briefly discuss this in Appendix H where we show how state regression outperforms weak derivative regression when the measurement sampling frequency is low. That being said, there exists a suite of methods from the state estimation / data assimilation fields for estimating $\mathbf{x}(t)$ given an uncertain measurement and dynamics model. We have mentioned this possible extension in our newly added "future work" section.