We first would like to thank the reviewers for their insightful comments and suggestions. We will give a list of changes that we propose to implement to improve the quality of the paper, based on the four reviews.

Foreword: This paper is framed as a methodological and theoretical contribution, with simple experimental validation. There is in particular no specific ethical concern with this paper – something we will add in a "Broader Impact" section.

Reviewer 1 • *Corollary 1*: Continuity and the permutation invariance conditions are satisfied in both the linear system and the AC power flow problems. The uniqueness is only proven for the linear system. However it is very likely to hold for many problems. We will add one sentence along this line. • *Correlation with 'ground truth'*: RMSE and correlation (between DSS and 'exact' solutions) provide complementary views. In this context we used correlation, we will add RMSE results in supplementary material. • *AC power flow discussion*: Indeed, a large loss does not necessarily imply poor predictions in V and θ, nor, in turn, in P_{ij} and Q_{ij}. But the approach remains beneficial in terms of computational cost as long as the predictions for P_{ij} and Q_{ij} are good enough. This will be better explained in the final version. We will also modify the caption of Table 2. • *"DGM" reference*: We will add it to references [6]-[7].

Reviewer 2 • *Critical applications*: Our primary application domain being the highly sensitive area of power grids, we are truly aware that AI can bring novel threats to systems security and reliability. This was our primary motivation for properly introducing mathematical concepts and ideas in this work. • *Limitations*: We are aware that the universal approximation theorem that we derive does not offer any guarantee of convergence. We will revise the paper to stress this more clearly. However, this non-trivial result is a pre-requisite to provide solid theoretical ground to the proposed approach. • *Experimental validation*: Adding more difficult use cases (pushing the loads to the limit in the power flow, handling non-homogeneous and even nonlinear PDEs as suggested, etc.) is on-going work. We will more clearly discuss the current limitations in the conclusion. In particular, we will mention the fact that in the case of critical applications, our approach can be used to find a good starting point for some classical method, thus saving large amounts of computing resources. • *CPU time*: We probably did not present our results clearly enough, but the unit used for CPU times in Table 1 is ms, making our estimation in line with the one from the review. We will modify the row and column captions. • *Adaptive number of iterations*: During our preliminary experiments we observed that having different neural network blocks at each propagation step gave better results, thus making $\overline{k}$ a fixed hyperparameter. However, a recurrent graph neural network structure with adaptive $\overline{k}$ is definitely an interesting avenue to explore. • *AC linear domain*: This aspect has been further investigated in some of our previous work, which we will cite in the revised version, by comparing our statistical solver, the Newton-Raphson method, and the DC-approximation (which relies on a linearization of the Kirchhoff's equations). These experiments showed that the DC-approximation was significantly worse than our proposed method, thus proving that the DSS performs well in the non linear domain of the AC PF problem. We acknowledge that this aspect should have been addressed in the paper. • *AC feasibility*: Our architecture takes as input the active and reactive loads, active production and voltage setpoints at generators, and we use a slack bus so the AC feasibility is not an issue. • *References*: We were not aware of the Shental et al. reference, and will add it in the discussion. • *Amount of samples*: We do need some sample problems, but we do not need to have their solutions, as in the supervised "proxy" approach, since we have a closed-form expression for the loss function. We will stress this better in the revised version. Nevertheless, it is true that the choice of the problems (i.e. of the distribution $\mathcal{D}$) is another important issue regarding the validation. However, the same issue arises in the proxy case. • *Notations*: $\mathbb{R}^{d_A}$ and $\mathbb{R}^{d_B}$ are introduced on line 138 and M_θ^k on line line 137, before being used in eq. (8). We will regroup clear notation definitions in the revised version for additional clarity. • *Hypotheses*: In addition to the continuity and permutation invariance hypotheses, we also need to have some uniqueness property as introduced in Section III to be able to properly lay the ground for the theorem. • *Choice of $\overline{k}$ and d*: If $\mathcal{D}$ is a dataset of snapshots of the Californian power grid, then one can compute the power grid diameter, which we proved to be a good lower bound for $\overline{k}$. Moreover, one can expect that the larger the inputs d_A or d_B, the larger d should be. We will add a sentence in the revised version. • *Hyperparameters*: We will add the ranges of the grid search for all hyperparameters.

Reviewer 3 • *Labelled data and 'proxy' approach*: We use the word 'proxy' to designate the method that learns to reproduce existing solutions. Here, we only need some sample problems, but not their solutions, thanks to the closed-form expression for the loss function. We will stress, and phrase, this better in the revised version. • *Correlation with 'ground truth'*: See response to reviewer 1. • *Computational complexity*: The theoretical complexity has been addressed line 152. The wall-clock times have been reported in Table 1 and Table 2 for all considered methods. • *Meaning of* $\mathbf{A}$ *and* $\mathbf{B}$: A and B are the matrix (resp. vector) that result from the discretization and assembling process of the Poisson equation. The chosen loss amounts to solve $AU = B$. Each element of U corresponds to a node that is shown in Figure (4). Moreover, the edges that are displayed at the top left correspond to non zero values A_{ij}. Further details are provided in lines 75-77 and Appendix A. • *Phrasing*: The term "brick" is employed to express the idea that neural networks are here elementary building blocks in a broader architecture. We will modify the wording.

Reviewer 4 • *Constraints*: The term "constraint" is used in the introduction: it does not refer to optimization constraints but to domain specific consideration. As defined in eq. (2)-(3), we focus on unconstrained optimization. Adding non-trivial constraints is actually a line of work we are currently investigating. We will correct this misleading phrasing. • *Comparison with proxy solutions*: We fully agree regarding the lack of comparison with proxy neural networks. This is on-going work.