We thank the reviewers for their valuable and positive feedback. We will address their main concerns:

Reviewer #1. We thank the reviewer for the positive and encouraging feedback and address their suggestions and questions: *1) On Clarity* ***"TPR and FDR same Figure":*** For space restriction, and in line with prior papers eg HRT [8], we elected to have them in the same plot, but we will make sure to remind the reader of the expected trends in the legend of the figure. *2) Comparison to the* ***Gate Formulation in (P):*** In our early experiments using a NN critic, we found this formulation to be under-performing. Theoretically, Formulation (P) does not lead to a convex formulation in the RKHS. (P) was indeed also studied in the cited reference [10] (Appendix A), using the Hilbert-Schmidt Independence Criterion and a concave approximation of ℓ_0, and was found to be under-performing as well. *3)* ***"Analysis of the Consistency":*** This is an interesting question. In order to recover the correct conditional independence we elected to use FDR control techniques to perform those dependent hypotheses testing (btw coordinates). By combining SIC with HRT and knockoffs we can guarantee that the correct dependency is recovered while the FDR is under control. For the consistency of SIC in the classical sense, one needs to analyze the solution of SIC, when the critic is not constrained to belong to an RKHS. This can be done by studying the solution of the equivalent PDE corresponding to this problem (which is challenging, but we think it can be also managed through the $\eta-$ trick). Then one would proceed by finding 1) conditions under which this solution exists in the RKHS 2) Generalization bounds from samples to the population solution in the RKHS. We can mention these points in the final version, but we feel that a thorough analysis is beyond the scope of the paper and will be left for future work. *4)* ***"ReLu, No Biases":*** The choice of Relu with no biases is dictated by requiring interpretability in terms of "input sparsity" of the neural network. In practice, there is no restriction on the network and one can use Relu with biases if input sparsity is not a goal. In any case, removing biases in the ReLU network does not seem to affect performance strongly, especially when the input data is centered.

Reviewer #2. We thank the reviewer for the positive and encouraging feedback and address their main concerns here. **Additional Comparison on the real word datasets:** For the CCLE experiments, we compared to ElasticNet which is the established feature selection method on this dataset [8, 36]. For the HIV knockoff experiments, we compared to the GLM feature selection of Candes et al [9]. We plan to include the Random Forest feature selections in the final version as an additional baseline, as suggested by the reviewer.

Reviewer #3. We thank the reviewer for the review and address their questions: *1)* ***HRT versus Knockoffs:*** Thank you for pointing this out and allowing us to clarify that this paper is not a comparative study between the two methods, rather it is a strength of SIC that it can be used with both. For a comparison between HRT and knockoffs, we refer the reviewer to [8], which shows similar performance for either method in terms of controlling FDR. We will highlight in the final paper that each method has its advantages. In HRT most of the computation is in 1) training the generative models, and 2) performing the randomization test, i.e. forwarding the data through the critic and computing p-values for each coordinate for R runs. On the other hand, if knockoff features can be modelled as multivariate Gaussians, controlling FDR with knockoffs can be done very cheaply, since it does not require randomization tests. If instead knockoff features have to be generated through non-linear models, knockoffs can be computationally expensive as well (for example Deep knockoffs, Romano et al. ICLR 2019). *2)* ***Kernel versus NN SIC:*** In early experiments, we tried using random networks defined by random Fourier features (approximating a Gaussian kernel), and we found that a fully trained network outperforms the fixed network. *3)* ***Computational Complexity versus Performance:*** We thank the reviewer for the question, and plan to clarify this point as follows. The cost of training SIC with SGD and mirror descent is of the same order of magnitude as training the base regressor neural network via back-propagation. The only additional overhead is the gradient penalty, where the cost is for a double back-propagation. In our experiments, this added computational cost is not an issue when training is performed on GPU. *4)* ***Conditions on*** Φ***:*** In our framework, Φ is defined on $\mathcal{X} \times \mathcal{Y}$, but is not restricted to be an outer product between feature maps. It is possible to ensure that $SIC(p_{xy}, p_x p_y) = 0$ iff $p_{xy} = p_x p_y$ by imposing universality on the feature map Φ to get injectivity of the mean embeddings. However, studying the conditions on the kernels under which $\eta_j = 0$ leads to $x_j \perp y | x_{-j}$ is an interesting open question, but beyond the scope of this paper. Instead, we tackled the problem (whether η_j are statistically significantly above zero) indirectly using HRT and Knockoffs that have theoretical guarantees to control the FDR with no restriction on the feature selection method used. *5)* ***Compact level sets of*** L_ϵ***:*** This condition is met when perturbing the problem with ϵ. This in not needed for the proof of Theorem 1, but is needed to guarantee that alternating optimization or first order methods on u and η are convergent (See the monograph [19] page 59). We address the minor comments: ***1) Generalized MI.*** There are many generalizations of mutual information (MI) such as the Renyi MI that uses Renyi divergence and many other extensions have been developed. For an introduction on this topic we refer the reviewer to α **mutual information** by Sergio Verdu. ***2-3) "zero norm of*** w***,*** $\frac{1}{N}$***"*** ℓ_0 norm and $\frac{1}{N}$ should be there. Thanks, typos will be corrected. ***4) Multiple runs? GLM?:*** This is for a single run, following the same experimental protocol as Candes et al. [9]. GLM stands for Generalized Linear Model. Boosted SIC is an ensemble of η for the same data, different random seeds initialization of the NN. **"**L **versus** g**"** In the cost L the only non-obvious terms for proving convexity in η and u are of the form $g(u, \eta_j) = \frac{u^\top A_j u}{\eta_j}$ (A_j PD) hence the proof is here only for those terms.