We thank all reviewers for all the constructive comments. We are encouraged that they found our contribution novel and significant toward feature selection and explainable AI (R4). We are glad they found the methods rigorously presented with clear and formal definitions (R1), the theoretical grounding was clear and easy to follow, and also they found that the approach is general enough to be of wide interest to the NeurIPS community (R2). Moreover we are pleased that they found our result on gene expression important (R2,R3). We address reviewers minor comments below and will incorporate all feedback in our revised version.

[R1] **Complex architecture, non-convex optimization, lack of guarantee in solutions obtained in practice (local minima etc).** We maximize the variational lower bound of Mutual Information (MI) as a surrogate to maximizing MI itself. The approximate distributions are modeled using a flexible neural network and optimized via stochastic gradient descent (SGD). Therefore, while the objective is non-convex, our solution is still guaranteed to reach the local minimum. [R1] **Time Complexity**: Since most competing deep neural net methods also used SGD, our complexity are comparable to these deep models and scalable to sample size. We have also observed a comparable run-time experimentally, but choose to highlight the interpretability experiments. We'll add run-time in the appendix.

[R2] **Gene expression Fig. 5 and 6 results.** As described by the captions of Fig. 5 and 6, they are not figures of the gene expression, nor is it a correlation matrix. Fig. 5 displays an indicator matrix (white selected; black not) telling us which genes *work together* to best predict the smoking status for each patient. Based on Def. 1 and 2, we discovered the features that are most dependent on each other and the label. Note that Fig. 5 indeed shows that the genes in the first top rows are selected by non-smokers and the genes in the lower rows are selected by smokers. Note that this cannot be discovered by correlation feature clustering alone because it gives the same cluster of features (groups) for all samples. [R2] **Direct ref for the Gumbel Softmax trick?** Yes, we have cited Ref. [22] and provided a background review in the paper for completeness. [R2]**Other Methods to Compare.** Note that we have compared our method to *nine* competing methods from various perspectives to feature selection (FS): global FS, deep instance-wise FS, deep unsupervised FS, and global FS with feature group learning (includes feature group clustering). We can add results on attention models to Fig. 7 in the appendix. They provide focus areas in the image rather than all the similarly important features (pixels).

[R3] **Thms 1 and 2. $I(\hat{\mathbf{X}};\mathbf{X})$ is upper bounded by the $H(X)$. Is the upper bound reachable?** Theorem 1 and 2 states that $I(\hat{\mathbf{X}};\mathbf{X}) = I(\mathbf{X};\mathbf{X}) \iff I(\mathbf{X};\mathbf{X}|\mathbf{Z}) = 0$ and $I(\bar{\mathbf{X}};\mathbf{Y}) = I(\mathbf{X};\mathbf{Y}) \iff I(\mathbf{X};\mathbf{Y}|\mathbf{Z}\odot \mathrm{s}) = 0$ respectively . This statement is not a guarantee of global optimum for LHS (i.e, $I(\hat{\mathbf{X}};\mathbf{X})$), but rather a justification of why maximizing LHS can lead to desirable redundancies in RHS (i.e. $I(\mathbf{X};\mathbf{X}|\mathbf{Z}) = 0$). By using neural network, we have experimentally shown (for genetic and MNIST) that it is sufficiently flexible to achieve highly accurate results. [R3]**Theorem 3 might not be useful.** We use Thm 3 to reveal how our definition of redundancy can be used to reveal the appropriate number of feature groups (a difficult issue in any clustering problem). It also supplies the theoretical background for us to discuss Thm 4, where the required assumption matches many genetic and public datasets. [R3] **Lasso Performing Better.** For the 2 cases in synthetic dataset where the correlation pattern was global, Lasso performed better; but when there is a mixture of correlation patterns their accuracy decreased to 74 percent. In general, such as in biology and image, data often have wide variations even within the same class. Our method is able to capture both patterns at a cost of a slight accuracy degradation (99.7% and 95% acc) on simple models. [R3] **Mutual Information (MI) Lower Bound.** We show in App. G that the objective $\max_{\theta_G,\theta_S} I(\mathbf{Y};\bar{\mathbf{X}})$ can be alternatively maximized with $\max_{\theta_G,\theta_S} E_{\mathbf{Y},\bar{\mathbf{X}}}[\log P(\mathbf{Y}|\bar{\mathbf{X}})]$ via Monte Carlo estimation, drawing samples from $P(\mathbf{Y},\bar{\mathbf{X}}) = P(\mathbf{Y}|\bar{\mathbf{X}})P(\bar{\mathbf{X}})$. This requires us to obtain samples from $\bar{\mathbf{X}}$ and know the distribution $P(\mathbf{Y}|\bar{\mathbf{X}})$. Since $P(\mathbf{Y}|\bar{\mathbf{X}})$ is not known, we then approximate this distribution via a neural network Q_{θ_P}. Eq. (8) is necessary because it provides the ancestral sampling steps to obtain samples for $\bar{\mathbf{X}}$. A similar logic then follows for Eq. (7) and the distribution $P(\mathbf{X},\hat{\mathbf{X}})$. We will adjust this section to better explain the process.

[R3, R4]**Literature review on MI.** L2X, that we cite and learns instance-wise feature selection (FS), is based on MI. Many traditional (global) FS utilizes MI as criterion for selection(as it is a natural criterion for measuring dependency among random variables), including mRMR. mRMR maximizes feature relevance while minimizing feature redundancy to find the *global minimal subset of features*. In contrast, our method, learns *instance-wise group FS*. We differ in two ways: (1) In mRMR, if F_1 and F_2 are highly dependent, it'll only pick one of them if they are relevant to prediction. If F_1 and F_2 are relevant to prediction, we select both as a group (highlighting to domain scientists that these two features are both relevant and redundant to each other). (2) mRMR is global – all samples select the same features; our method is instance-wise – we provide which feature groups are important to each sample. Thank you for the advice; we'll include mRMR and MI-based FS in the revision. [R4] **How to stop each instance obtaining group of features and overwhelm users in terms of cognitive attention:** All the groups are being learned through a neural network which is a continuous mapping function, hence if two samples are very similar, the model would give a similar representation for the groups that are important for the prediction.