We are truly appreciative to all reviewers for their insightful and very helpful comments. Overall, we propose a novel mutual information (MI) regularisation method to remove continuous target-information from latent representations. We believe our work should be shared with the community as it demonstrates the effectiveness of our method and has numerous socially-relevant applications such as drug discovery and solar cell design. We emphasise that moving from discrete to continuous targets is not straightforward as minimising MI in such settings presents several difficulties. Thus, we view our method as a novel solution to a challenging problem. Moreover, the reviewers acknowledged that the paper is "well written and easy to follow" (R1, R8), "very useful in molecular design" (R2), "novel and interesting" (R1, R2, R4, R3), "well motivated" (R4,R8) and the "method appears sound" (R1,R5). We address the reviewer comments below:

(R1, R2) **Claim ll. 274–276, what influence Z_0 actually has on generated structures?, Paper strengthened by sampling from generative model.** We agree that this is a highly important and relevant question. Firstly, it is not possible to capture all information in the 2-dim. Z_1 since we require at least 16 dim. for a good reconstruction (see Appendix, Exp. 2 + Fig. 3) Secondly, we performed generative experiments (see Add. Exp. 4) on the highly complex Zinc dataset (>250k molecules) to demonstrate the effectiveness of our approach. We will add the same experiments for QM9 to the appendix in the final version.

(R2) **The Kraskov estimator can perform inferior when MI is high, Estimate only lower bound.** This remark is true, however, in the regime we consider (minimising the MI), there is almost no dependence between Z_0 and Y. In addition, we demonstrated that the MI is small by looking at the qualitative results of the latent space (e.g. Fig. 4) and at the generative nature of our model (Appendix, Add. Exp. 4).

(R2) **is Z_0 even necessary for reconstruction now?, Reconstructed samples are only supported in 1D and thus could be completely explained by Z_1 alone?** This is a misunderstanding. The dataset is constructed such that every point on the diagonal (X) maps to the same point in Y. Therefore, we need additional dimensions to reconstruct the position on the diagonal. We performed an additional experiment where Z is only 1D. This leads to a MAE(X)=1.97 and a MAE(Y)=0.67 which indicates that a 1D space is not sufficient to reconstruct X.

(R2, R4, R5) **paper aims to study symmetric transformation learning, but it mostly talks about disentangled representation learning, unclear what "symmetry" means here, When considering symmetries one usually has certain geometric operations in mind (such as the rotation cited in Fig. 1a).** As we state in the introduction (line 33), the goal of the paper is to learn a symmetry property f of the system that leads to a predefined invariance (Y). The purpose of our model, however, is to go beyond simple geometric operations and to allow for learning arbitrary continuous transformations that result in the invariance. In general, by considering arbitrary continuous transformations g (Fig. 1), we model the group action of a Lie group (the set of g) on the space X that preserves the symmetry f. We will also extend the related work section with a survey of related disentanglement approaches.

(R2, R4, R5) **It it is absolutely unclear why h, h^{-1} should form a bijection, there should be more explanation about the technique of relaxing Gaussian assumption with bijective mapping, The bijection employed in Fig. 3 is essentially an invertible network.** Since both h and h^{-1} are functions between continuous sets, the loss given in line 181 can only be 0 if both functions form a one-to-one mapping. Thus, Eqs. (8,9) do measure the actual mutual information. This is indeed also a feature of an invertible network and using one is a valid alternative the relaxation technique we employed.

(R4) **In the submitted codes, you calculate the bijective loss with $||h^{-1}(h(Y)) - h(Y)||$, which is different from what you defined in the paper $||h^{-1}(h(Y)) - Y||$.** The correct equation is $||h^{-1}(h(Y)) - Y||$. We uploaded the wrong code, the corrected results with the loss in the paper are: MAE(X)=0.05, MAE(Y)=0.44, MI(Z0,Y)=0.19. The results in the real experiments are not affected as the property data is approx. Gaussian which is why we have not used the bijection extension.

(R3, R4) **Why symmetry is very important? What is the benefit of the method for chemistry?**. In material science, e.g. solar cell design, we want to find all variations of molecules that posses the same bandgap energy of 1.2 eV to adequately generate electricity. Therefore, we need to find a transformation that alters a molecule and leaves the property unchanged (see ll. 22–32).

(R3, R4) **this work is a slightly-modified version of GAN+VAE framework. Please illustrate more insightful contents or the major differences.** Moving from discrete to continuous targets is not straightforward as minimising MI in such settings gives rise to several difficulties. To the best of our knowledge, cognate models have solely focused on discrete Y. This is because naively using the negative log-likelihood (NLL) as done when maximising mutual information in other deep information bottleneck models leads to critical problems in continuous domains. This stems from the fact that fundamental properties of mutual information, such as invariance to one-to-one transformations, are not captured by this mutual information estimator. Moreover, we want to consider multiple properties at once, where every one requires high resolution. Simultaneous high-resolutional discretisation of multiple targets would result in an intractable classification problem.

(R2) **discussion of the difficulty of estimating MI in the experiments is not given.** Throughout our model, we use the analytic formula for Gaussian MI (Eqs. (8,9)) which we extend with the Gaussian relaxation. We subsequently use the Kraskov estimator as a benchmark. A comparison to different approaches of MI estimation such as MINE is not a focal point of the paper, but we will add a short discussion of suggested related methods in case of acceptance.