On the Generative Utility of Cyclic Conditionals

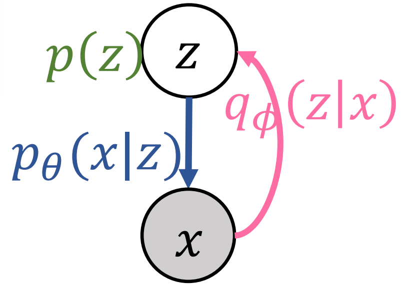
Chang Liu

Microsoft Research Asia

Joint work with:
Haoyue Tang, Tao Qin, Jintao Wang, Tie-Yan Liu.

Introduction

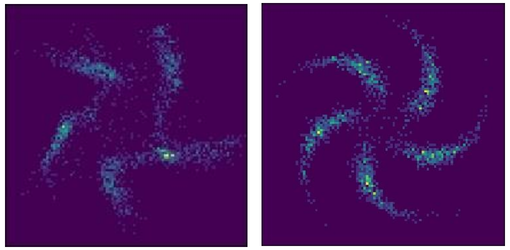
The problem:
Whether or when can we model a *joint* distribution $p(x, z)$
only using two *conditional* models $p(x|z)$ and $q(z|x)$ that form a cycle?



- Motivation from deep generative models:

  Use both a $p(x|z)$ model for *generation*, and a $q(z|x)$ model for *representation*.
  But define their common joint distr. by a prior $p(z)$: $p(x, z) \coloneqq p(z)p(x|z)$.

  - Standard Gaussian prior:
    - **Manifold mismatch**:
      $p(x)$ has a *simply connected support* as $p(z)$ $\Longrightarrow$ restricted expressiveness.



    - **Posterior collapse**:
      $q(z|x)$ is squeezed to the origin $\Longrightarrow$ degraded representativeness.

  - Using an informative prior:

    Domain knowledge on the prior is less common than on the conditional models.
    (e.g., shift/rotation invariance of $q(z|x)$ for image representation (CNN/SphereNet))
  - Learning a prior model: additional cost; harder than learning conditional models.

Introduction

The problem:
Whether or when can we model a *joint* distribution $p(x, z)$
only using two *conditional* models $p(x|z)$ and $q(z|x)$ that form a cycle?

- Key sub-problems:
  - **Compatibility** (existence): When the two conditionals can be induced from a common joint.
  - **Determinacy** (uniqueness): When the two *compatible* conditionals uniquely determine a joint.
- In this work,
  - Theory: **compatibility** criteria (equivalent conditions) and sufficient conditions for **determinacy**.
    - Operable and self-contained.
    - Unify continuous and discrete cases.
  - **CyGen**: **Cy**clic-conditional **Gen**erative model.
    - Methods for enforcing compatibility, fitting data, and data generation.

Related Work: Modeling

- Cyclic conditional models
  - Dependency networks [Heckerman’00]:
    No latent variable (so compatibility is not a problem). Gibbs sampling for the joint.
  - Denoising auto-encoders (DAEs) [Vincent’08]: $\min \mathbb{E}_{p^*(x) q(z|x)}[\log p(x|z)]$.
    - Variants: Uncertainty AE [Grover’19], Walkback [Bengio’13], GibbsNet [Lamb’17].
    - The loss is not suitable for optimizing $q(z|x)$ (mode-collapse, weakens determinacy).
    - Inefficient generation and unstable training by Gibbs sampling.
  - Dual learning [He’16; Xia’17a,b; Lin’19], Disco[Kim’17]/Cycle[Zhu’17]/Dual[Yi’17]-GAN:
    - Not for generative modeling (in fact, they lack determinacy).
    - No latent variable, unpaired data.

Related Work: Theory

- Compatibility
  - The classical condition [Arnold'89,01] is not necessary.
  - The equivalent condition [Berti'14] is still existential.
  - Results from DAE [Bengio'13,14; Lamb'17; Grover'19]: not self-contained ($p^*(x)$ is required).
  - Cycle-consistency loss [Kim'17; Zhu'17; Yi'17; Lin'19]: only for deterministic conditionals.
- Determinacy
  - Determining $p(x)$ through score matching (SM):
    DAE $\Leftrightarrow$ denoising SM (Gaussian RBM) [Vincent'11].
    DAE $\Leftrightarrow$ SM (Gaussian decoder noise and infinitesimal Gaussian corruption) [Alain'14].
  - Determining $p(x, z)$ through Gibbs chain:
    - The chain is ergodic thus has a unique stationary distr. $\pi(x, z)$ under a global [Bengio'13; Lamb'17; Grover'19] or local [Bengio'13] shared support condition.
    - When incompatible, $\pi(z|x) \neq q(z|x)$ or $\pi(x|z) \neq p(x|z)$ (depending on the order of vars.) [Heckerman'00, Bengio'13].
    - No explicit expression for learning. Slow convergence for data generation (and learning for Walkback and GibbsNet).

Theory

Setup

- Measure spaces for random variables x and z: $(\mathbb{X}, \mathscr{X}, \xi)$ and $(\mathbb{Z}, \mathscr{Z}, \zeta)$.
- Product measure space $(\mathbb{X} \times \mathbb{Z}, \mathscr{X} \bigotimes \mathscr{Z}, \xi \otimes \zeta)$.
- For $\mathcal{W} \in \mathscr{X} \bigotimes \mathscr{Z}$, define
  its *slice* at z: $\mathcal{W}_z \coloneqq \{x \mid (x, z) \in \mathcal{W}\}$,
  its *projection* onto $\mathbb{Z}$: $\mathcal{W}^{\mathbb{Z}} \coloneqq \{z \mid \exists x \text{ s.t.} (x, z) \in \mathcal{W}\}$.
- For a joint distribution π, define
  its *marginal* onto $\mathbb{Z}$: $\pi^{\mathbb{Z}}(\mathcal{Z}) \coloneqq \pi(\mathbb{X} \times \mathcal{Z})$,
  its *conditional* $\pi(\mathcal{X}|z) \coloneqq \frac{\mathrm{d}\pi(\mathcal{X} \times \cdot)}{\mathrm{d}\pi^{\mathbb{Z}}(\cdot)}(z)$ (this is **only $\boldsymbol{\pi}^{\mathbb{Z}}$-a.s. unique**).
- Define "$=^{\xi}$", "$\subseteq^{\xi}$" as the extensions of "$=$", "$\subseteq$" up to a set of ξ-measure-zero.

Theory

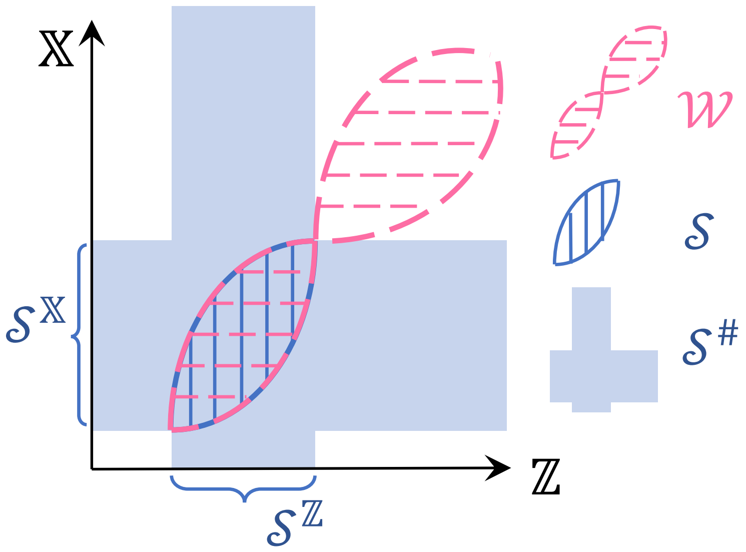
Absolutely continuous case

- For any z and x, $\mu(\cdot\,|z)$ and $\nu(\cdot\,|x)$ are either abs. cont. (w.r.t ξ and ζ), or a zero measure.
- Represented by density functions $p(x|z)$ and $q(z|x)$.
- Incl.: "smooth" distributions on Euclidean spaces / manifolds, and all distributions on finite/discrete spaces.
- Incl.: VAEs, diffusion-based models.

Theory

Absolutely continuous case

- Compatibility
  - First intuition: compatible $\Leftrightarrow$ the ratio $\frac{p(x|z)}{q(z|x)} = \frac{p(x,z)/p(z)}{p(x,z)/p(x)} = p(x)\frac{1}{p(z)}$ factorizes.
  - The classical condition [Arnold'89,01] requires the factorization over $\mathbb{X} \times \mathbb{Z}$:
    It is *not necessary*! Because $p(x|z)$ is uncontrolled outside the support of $\pi^{\mathbb{Z}}$.
    For identifying a proper region for the factorization,
  - **Definition**: A set $\mathcal{S}$ is said to be a $\xi \otimes \zeta$*-complete component* of $\mathcal{W} \in \mathscr{X} \otimes \mathscr{Z}$, **if** $\mathcal{S}^{\#} \cap \mathcal{W} =^{\xi \otimes \zeta} \mathcal{S}$, where $\mathcal{S}^{\#} := \mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{S}^{\mathbb{Z}}$ is the *stretch* of $\mathcal{S}$.
    - *Complete* under stretching and intersecting with $\mathcal{W}$: so that integration on $\mathcal{S}_z$ = integration on $\mathcal{W}_z$, for a.e. $z \in \mathcal{S}^{\mathbb{Z}}$. (similarly for $\mathcal{S}_x$, $\mathcal{W}_x$).
    - Conditionals are a.s. determined on $\mathcal{S}^{\#}$ if $\mathcal{S}$ is the support of the joint.



Theory

often just a few candidates, so it is *operable*.

$p(x|z)$ determines the distribution on $\mathbb{X} \times \{z\}$ if z is in the support, so $q(z|x)$ should respect it (>0 where $p(x|z)$ is) to avoid *support conflict*.

Absolutely continuous case

- **Theorem** (compatibility criterion, abs. cont.). $p(x|z)$ and $q(z|x)$ are compatible, **if and only if** there *exists* a set $\mathcal{S}$ (called *complete support*) such that:

makes conditionals *normalized*, since $\mathcal{S}_z =^{\xi} (\mathcal{W}_{p,q})_z = \mathcal{P}_z$.

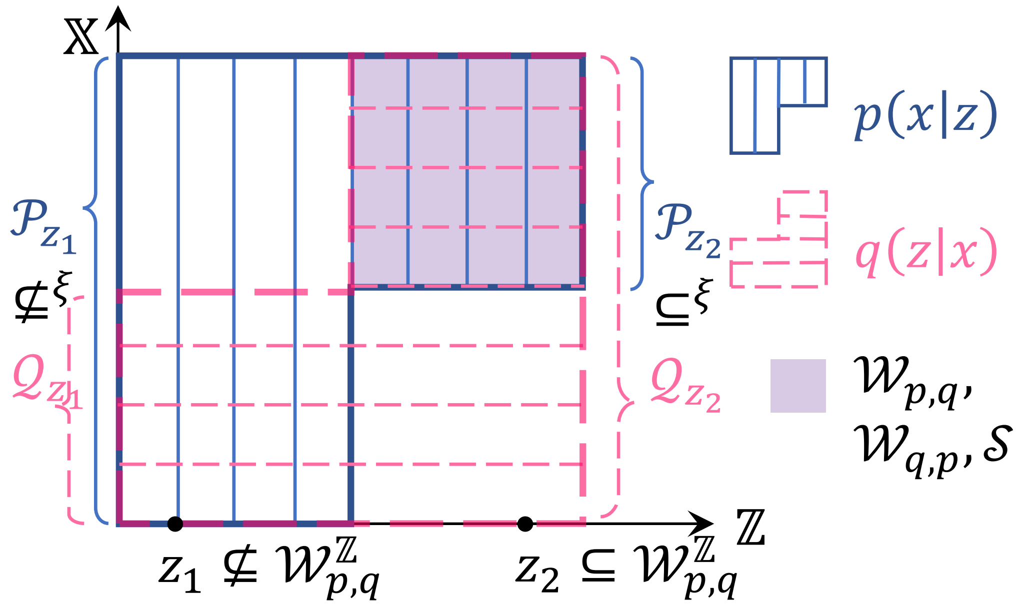
to make the ratio *well-defined*

for sufficiency; not guaranteed by (i)

the first intuition

(i) $\mathcal{S}$ is a $\xi \otimes \zeta$-*complete component* of both $\mathcal{W}_{p,q} := \bigcup_{z: \mathcal{P}_z \subseteq^{\xi} \mathcal{Q}_z} \mathcal{P}_z \times \{z\}$ and $\mathcal{W}_{q,p} := \bigcup_{x: \mathcal{Q}_x \subseteq^{\zeta} \mathcal{P}_x} \{x\} \times \mathcal{Q}_x$,
where $\mathcal{P}_z := \{x \mid p(x|z) > 0\}$, $\mathcal{P}_x := \{z \mid p(x|z) > 0\}$,
and $\mathcal{Q}_z := \{x \mid q(z|x) > 0\}$, $\mathcal{Q}_x := \{z \mid q(z|x) > 0\}$;
(ii) $\mathcal{S}^{\mathbb{X}} \subseteq^{\xi} \mathcal{W}_{q,p}^{\mathbb{X}}$, $\mathcal{S}^{\mathbb{Z}} \subseteq^{\zeta} \mathcal{W}_{p,q}^{\mathbb{Z}}$;
(iii) $(\xi \otimes \zeta)(\mathcal{S}) > 0$;
(iv) $\frac{p(x|z)}{q(z|x)}$ factorizes as $a(x)b(z)$, $\xi \otimes \zeta$-a.e. on $\mathcal{S}$;
(v) $a(x)$ is ξ-integrable on $\mathcal{S}^{\mathbb{X}}$.

For sufficiency, $\pi(\mathcal{W}) := \frac{\int_{\mathcal{W} \cap \mathcal{S}} q(z|x)|a(x)|(\xi \otimes \zeta)(\mathrm{d}x\mathrm{d}z)}{\int_{\mathcal{S}^{\mathbb{X}}} |a(x)|\xi(\mathrm{d}x)}$, $\forall \mathcal{W} \in \mathscr{X} \otimes \mathscr{Z}$ is a compatible joint.



Theory

Absolutely continuous case

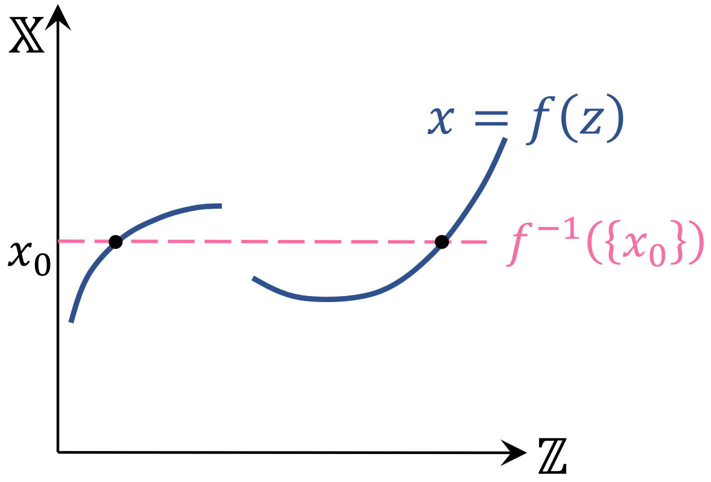
- **Theorem** (determinacy, abs. cont.). Let $\mathcal{S}$ be a complete support of *compatible* conditionals $p(x|z)$ and $q(z|x)$. **If** $\mathcal{S}_z =^{\xi} \mathcal{S}^{\mathbb{X}}$ for ζ-a.e. $z \in \mathcal{S}^{\mathbb{Z}}$ or $\mathcal{S}_x =^{\zeta} \mathcal{S}^{\mathbb{Z}}$ for ξ-a.e. $x \in \mathcal{S}^{\mathbb{X}}$, **then** their compatible joint supported on $\mathcal{S}$ is unique.
  - Roughly means $\mathcal{S}$ is “rectangular”: *irreducibility* of the Gibbs chain.
  - The uniqueness is only possible on each complete support $\mathcal{S}$.

- **Corollary**. **If** *compatible* conditionals $p(x|z)$ and $q(z|x)$ have a.e.-full supports, **then** their compatible joint on $\mathbb{X} \times \mathbb{Z}$ is unique.
  - Determinacy in the abs. cont. case is often *sufficient*.

Theory

Dirac case

- $\mu(\mathcal{X}|z) = \delta_{f(z)}(\mathcal{X}) \coloneqq \mathbb{I}[f(z) \in \mathcal{X}]$ ($f: \mathbb{Z} \to \mathbb{X}$ is measurable; e.g., when continuous).
- Incl.: Euclidean/manifold case (no density function), and finite/discrete case (*also abs. cont.*).
- Incl.: GANs, flow-based models.
- **Theorem** (compatibility criterion, Dirac). Suppose $\mathscr{X}$ contains all the single-point sets. Then conditional $\nu(\cdot|x)$ is compatible with $\mu(\mathcal{X}|z) = \delta_{f(z)}(\mathcal{X})$, **if and only if** there exists $x_0 \in \mathbb{X}$ s.t. $\nu(f^{-1}(\{x_0\})|x_0) = 1$.
  - $\nu(\cdot|x)$ is not required to concentrate on the curve for *any* x: for *one* such x_0, $\delta_{(x_0, f(x_0))}$ is already a compatible joint.



- When compatibility is desired on a set $\mathcal{X}$ and $\nu(\cdot|x) \coloneqq \delta_{g(x)}(\cdot)$:
  - $\min \mathbb{E}_{p(x)} \ell\Big(x, f\big(g(x)\big)\Big)$ is *sufficient* ($p(x)$ is supported on $\mathcal{X}$; ℓ is a metric): the **cycle-consistency loss** [Kim'17; Zhu'17; Yi'17; Lin'19].
  - It is also *necessary* if f is invertible: flow-based models are naturally compatible.

Theory

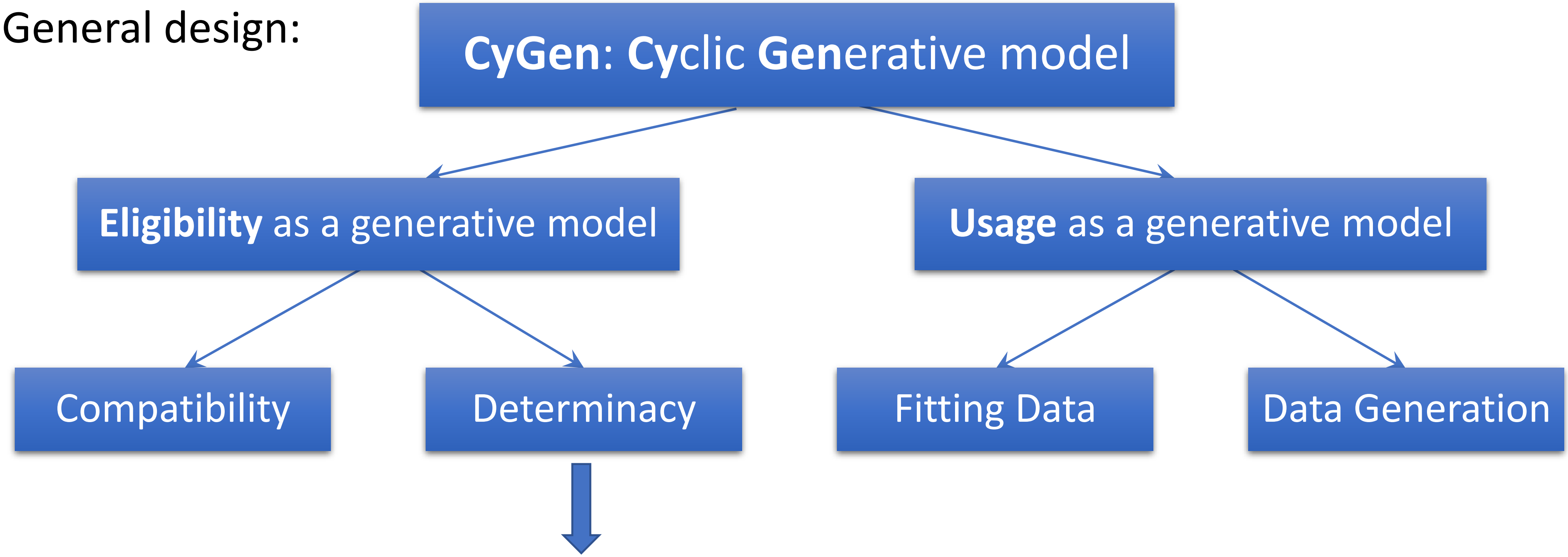
Dirac case

Determinacy:

- On each x_0 in the theorem, there is a compatible joint $\delta_{(x_0, f(x_0))}$.
- But if such an x_0 is not unique, the joint is *not unique* on $\mathbb{X} \times \mathbb{Z}$.
  - Determinacy in the Dirac case is usually *insufficient*:
    Compatible Dirac conditionals only determine a curve on $\mathbb{X} \times \mathbb{Z}$ but not a distribution on it.
  - If $f(z) \equiv x_0$ is constant, then the joint is fully determined by $\nu(\cdot \mid x_0)$.

CyGen

- General design:



- Use **abs. cont.** conditionals (like VAEs), since Dirac conditionals (like GANs, flow-based models) have *insufficient* determinacy.
- Modeled by parameterized **densities** $p_\theta(x|z)$, $q_\phi(z|x)$ with full supports.

CyGen

- Enforcing compatibility:

  $C(\theta,\phi) \coloneqq \mathbb{E}_{\rho(x,z)} \left\| \nabla_x \nabla_z^\top r_{\theta,\phi}(x,z) \right\|_F^2$, where $r_{\theta,\phi}(x,z) \coloneqq \log\left(p_\theta(x|z)/q_\phi(z|x)\right)$,

  and $\rho(x,z)$ is an abs. cont. reference distr. supported on $\mathbb{X} \times \mathbb{Z}$, e.g., $p^*(x) q_\phi(z|x)$.

  - $C(\theta,\phi) = 0 \Leftrightarrow p_\theta(x|z)/q_\phi(z|x)$ factorizes a.e.
  - Generalizes the *cycle-consistency loss* to *probabilistic* conditionals.
  - **Efficient implementation** by Hutchinson's ['89] trace estimator: $\mathrm{tr}(A) = \mathbb{E}_{p(\eta)}[\eta^\top A \eta]$

    ➔ $C(\theta,\phi) = \mathbb{E}_{\rho(x,z)} \mathbb{E}_{p(\eta)} \left\| \nabla_z \left(\eta^\top \nabla_x r_{\theta,\phi}(x,z) \right) \right\|_2^2$.

    #{derivative computation}: $O(d_{\mathbb{X}} d_{\mathbb{Z}})$ ➔ $O(d_{\mathbb{X}} + d_{\mathbb{Z}})$.

    $p(\eta)$ is any distr. s.t. $\mathbb{E}[\eta] = 0, \mathrm{Var}[\eta] = I$.
  - Gradient estimation for flows $q_\phi(z|x)$: $z = T_\phi(e|x), e \sim p(e)$ with *intractable inverse*:

    $\nabla_Z \log q_{Z|X}\left(T_\phi(e|x)|x\right) = \left(\nabla_e T_\phi^\top(e|x)\right)^{-1} \nabla_e h_\phi(e,x)$,

    $\nabla_X \log q_{Z|X}\left(T_\phi(e|x)|x\right) = \nabla_x h_\phi(e,x) - \left(\nabla_x T_\phi^\top(e|x)\right) \nabla_Z \log q_{Z|X}\left(T_\phi(e|x)|x\right)$,

    where $h_\phi(e,x) \coloneqq \log q_{Z|X}\left(T_\phi(e|x)|x\right)$.

CyGen

- Enforcing compatibility:

$$C(\theta, \phi) := \mathbb{E}_{\rho(x,z)} \left\| \nabla_x \nabla_z^\top r_{\theta,\phi}(x,z) \right\|_F^2, \text{ where } r_{\theta,\phi}(x,z) := \log\left(p_\theta(x|z) / q_\phi(z|x) \right).$$

  - Implication on Gaussian VAE $p_\theta(x|z) = \mathcal{N}\left(x \middle| f_\theta(z), \sigma_d^2 I\right)$, $q_\phi(z|x) = \mathcal{N}\left(z \middle| g_\phi(x), \sigma_e^2 I\right)$:

    $C(\theta, \phi) = \mathbb{E}_{\rho(x,z)} \left\| \frac{1}{\sigma_d^2} \left(\nabla_z f^\top(z)\right)^\top - \frac{1}{\sigma_e^2} \nabla_x g^\top(x) \right\|_F^2 = 0 \Longleftrightarrow f_\theta(z), g_\phi(x)$ are affine.

    - Meets conclusions in causality [Zhang'09; Peters'14].
    - Root cause of recent observation (latent space is quite linear [Shao'18]) and analysis (latent space coordinates the data manifold [Dai'19], encoder learns a rescaled isometric embedding [Nakagawa'21]).
    - For a nonlinear repr., use a more flexible $q_\phi(z|x)$ model (e.g., Sylvester flow [VDBerg'18]).
  - Relation to AE regularizations:
    - Contractive AE [Rifai'11]: $\mathbb{E}_{p^*(x)} \|\nabla g^\top(x)\|_F^2$.
    - Denoising AE [Rifai'11; Alain'14]: $\mathbb{E}_{p^*(x)} \|\nabla (f \circ g)^\top\|_F^2$ (Gauss. enc. noise, infinitesimal Gauss. corruption).
    - "Tied weights" in AEs [Vincent'08; Rifai'11; Alain'14]: compatibility for sigmoid conditionals.

CyGen

- Fitting data:
  - When compatible, $p_{\theta,\phi}(x) = 1/\int_{\mathbb{Z}} \frac{p_{\theta,\phi}(z')}{p_{\theta,\phi}(x)} \zeta(\mathrm{d}z') = 1/\mathbb{E}_{q_\phi(z'|x)}[1/p_\theta(x|z')]$.
  - Maximum Likelihood Estimate (MLE):
    $$\left(\min_{\theta,\phi}\right)\ \mathbb{E}_{p^*(x)}[-\log p_{\theta,\phi}(x)] = \mathbb{E}_{p^*(x)}\left[\log \mathbb{E}_{q_\phi(z'|x)}[1/p_\theta(x|z')]\right].$$
    - Estimate by \`logsumexp\`.
    - The DAE loss $\mathbb{E}_{p^*(x)q_\phi(z'|x)}[-\log p_\theta(x|z')]$
      - $\leq \mathbb{E}_{p^*(x)}[-\log p_{\theta,\phi}(x)]$: improper for MLE.
      - Makes $q_\phi(z'|x)$ mode-collapsed: hurts determinacy.
  - CyGen final training loss: $\left(\min_{\theta,\phi}\right)\ \mathbb{E}_{p^*(x)}[-\log p_{\theta,\phi}(x)] + \lambda\, C(\theta,\phi)$.

CyGen

- Data generation:

  Sample from the learned data distribution $p_{\theta,\phi}(x)$.

  - **Gibbs sampling**: iteratively sample from $p_\theta(x|z)$ and $q_\phi(z|x)$.
  - **Dynamics-based MCMC**:
    - Converges faster than Gibbs sampling.
    - Only needs an unnormalized density of $p_{\theta,\phi}(x)$, which is available:

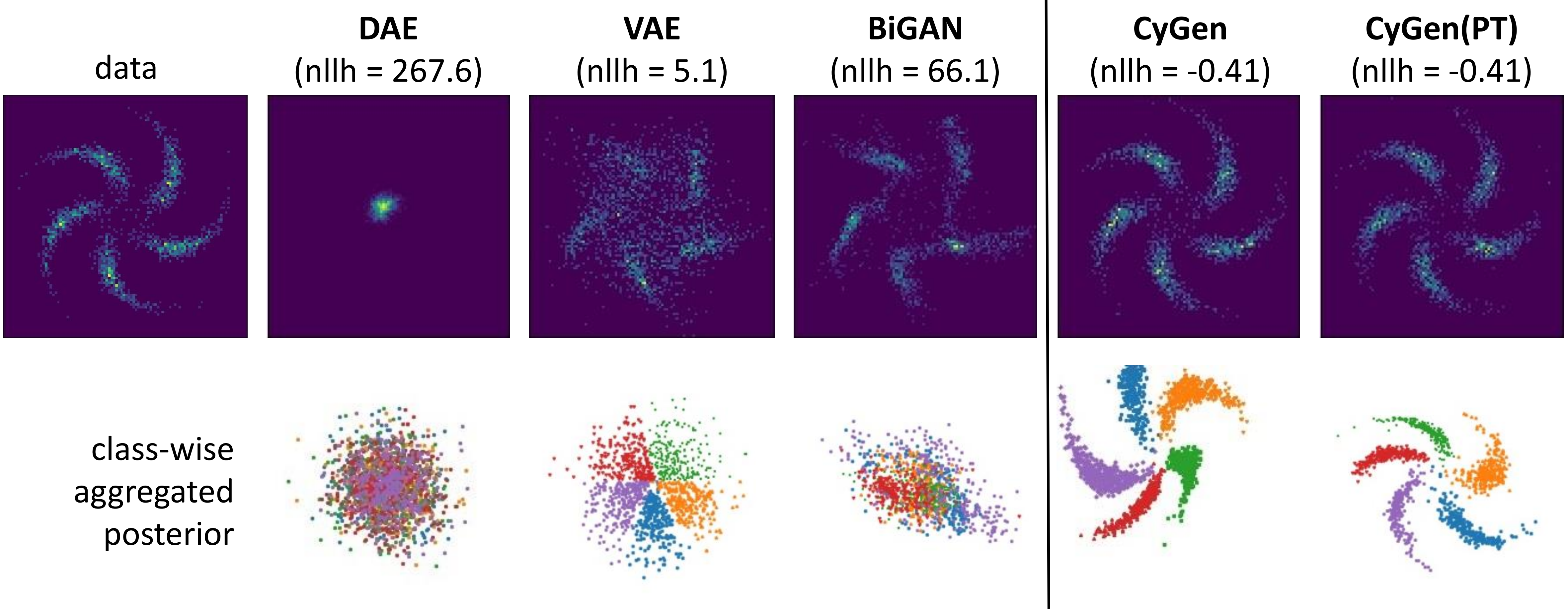
      $p_{\theta,\phi}(x) = \frac{p_{\theta,\phi}(x)}{p_{\theta,\phi}(z)} p_{\theta,\phi}(z) = \frac{p_\theta(x|z)}{q_\phi(z|x)} p_{\theta,\phi}(z) \propto \frac{p_\theta(x|z)}{q_\phi(z|x)}$, for any value of z.
    - E.g., Stochastic Gradient Langevin dynamics (SGLD):

      $x^{(t+1)} = x^{(t)} + \varepsilon \nabla_{x^{(t)}} \log \frac{p_\theta(x^{(t)}|z^{(t)})}{q_\phi(z^{(t)}|x^{(t)})} + \sqrt{2\varepsilon}\, \eta^{(t)}$, where $z^{(t)} \sim q_\phi(z|x^{(t)})$, $\eta^{(t)} \sim \mathcal{N}(0, I)$.

Experiment Results: Synthetic

Pre**T**rain as a VAE then mainly finetune $q_\phi(z|x)$.

- **Generation** and **Representation**: *manifold mismatch* and *posterior collapse* solved.



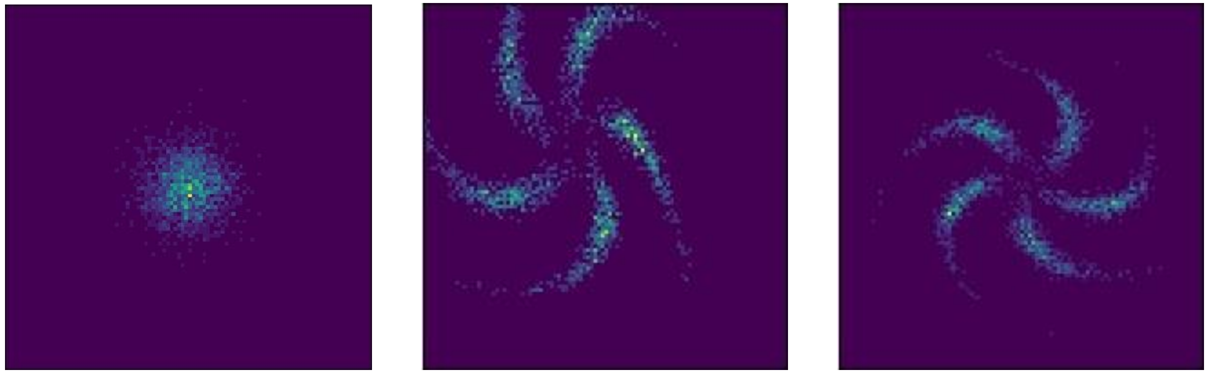
Experiment Results: Synthetic

- **Incorporating knowledge into conditional models**

  The VAE-pretrained $p_\theta(x|z)$ model encodes the knowledge:

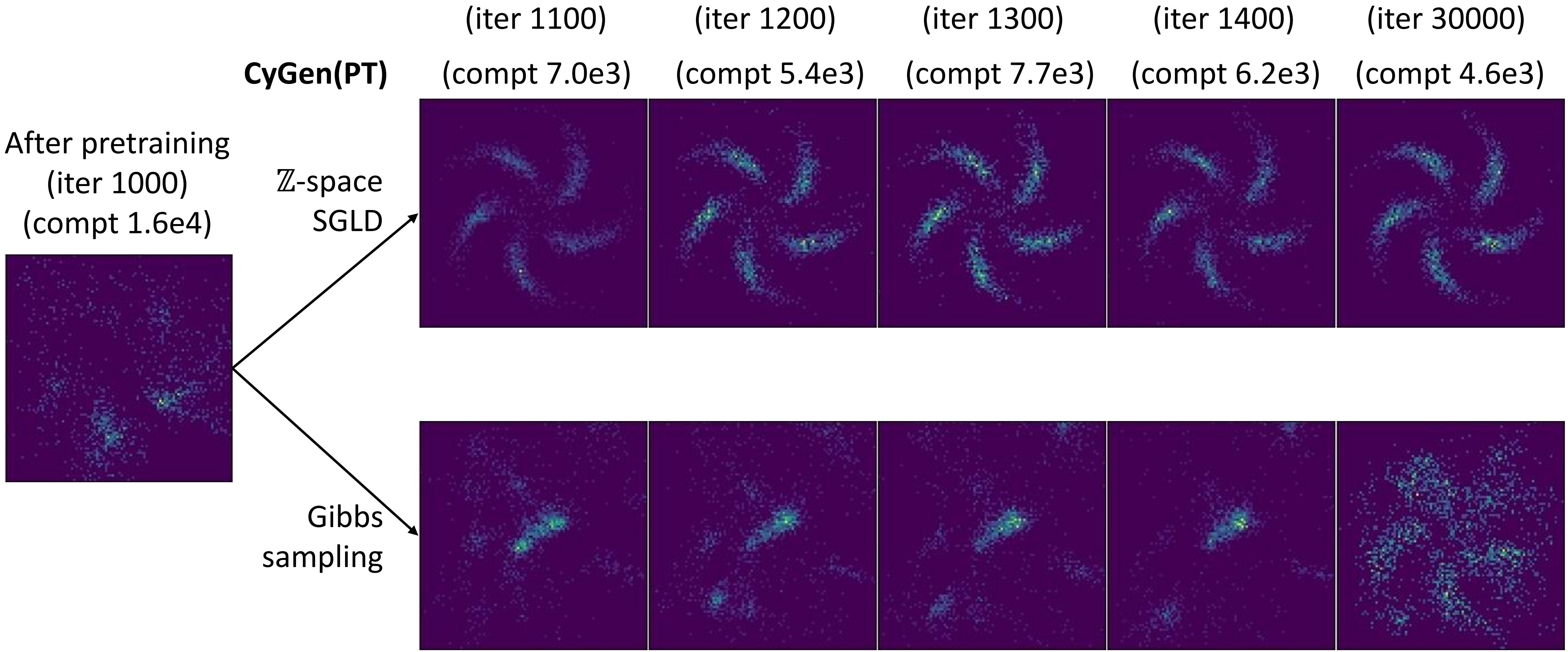
  "the prior is centered and centrosymmetric".

Prior of **VAE** | Prior of **CyGen** | Prior of **CyGen(PT)**



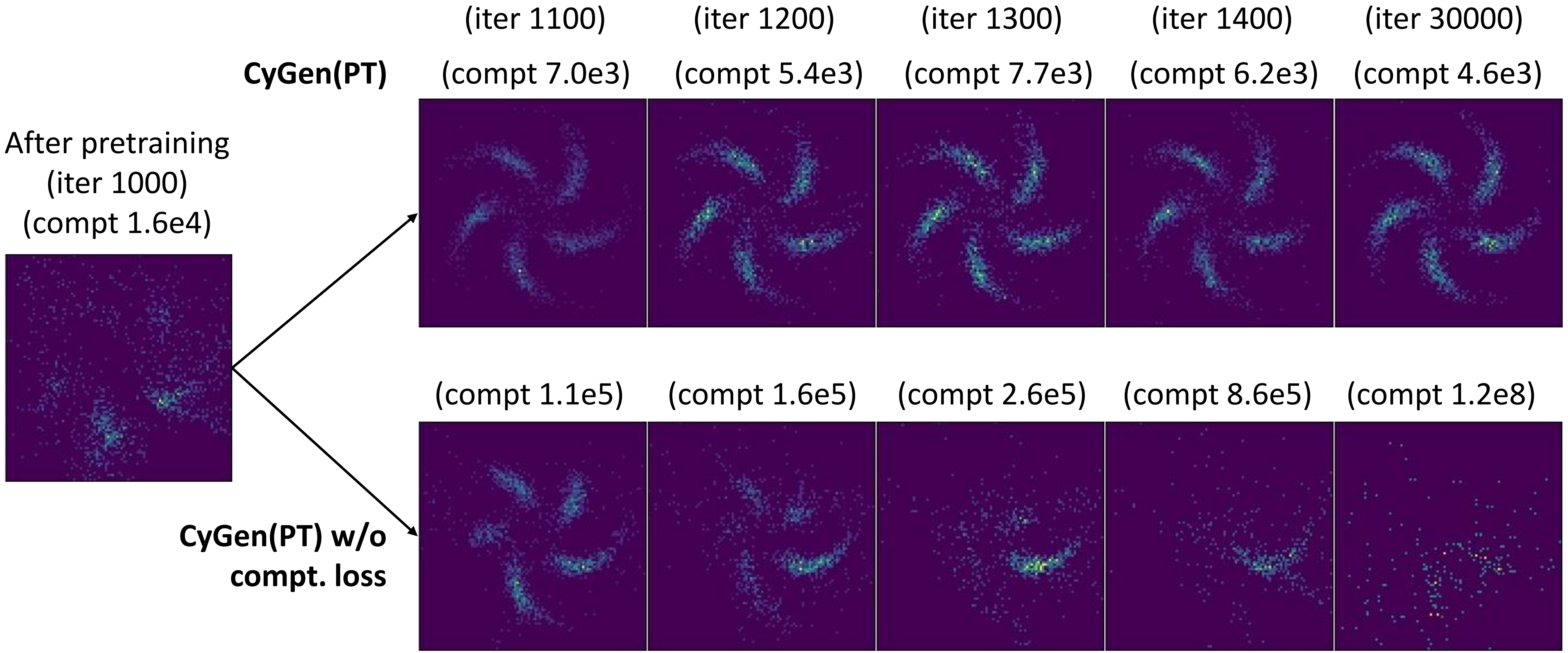
Experiment Results: Synthetic

- **Comparison of data generation methods**: SGLD is better and more robust to incompatibility.



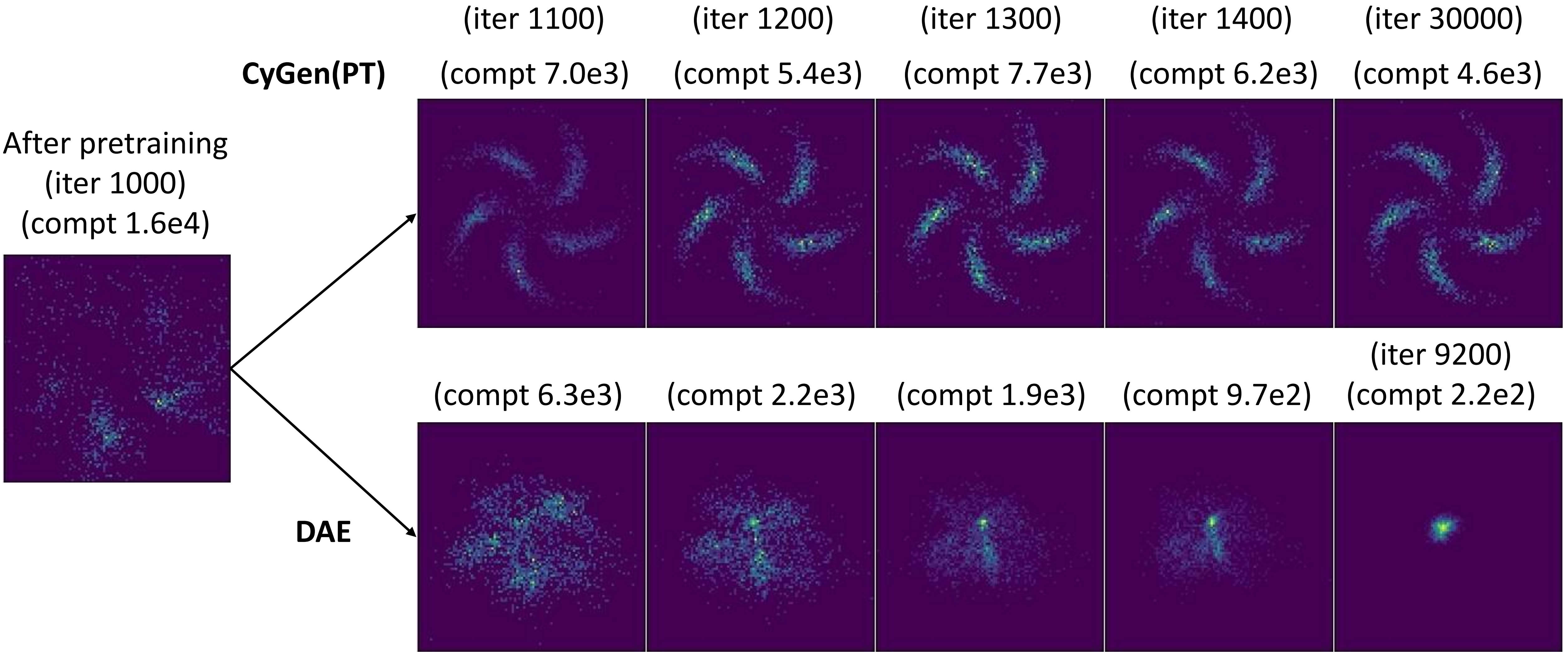
Experiment Results: Synthetic

- **Necessity of compatibility**



Experiment Results: Synthetic

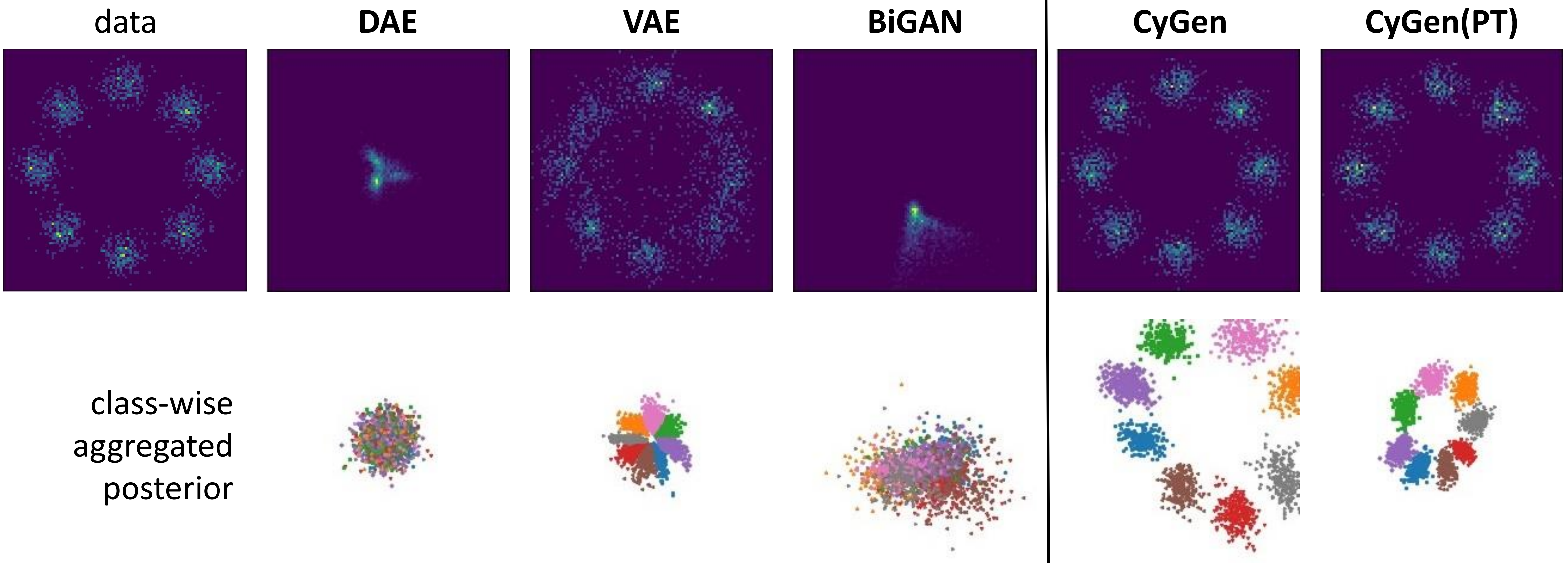
- **DAE mode collapse**



Experiment Results: Synthetic

- **Generation** and **Representation**: “8gaussians” dataset.

PreTrain as a VAE then mainly finetune $q_\phi(z|x)$.



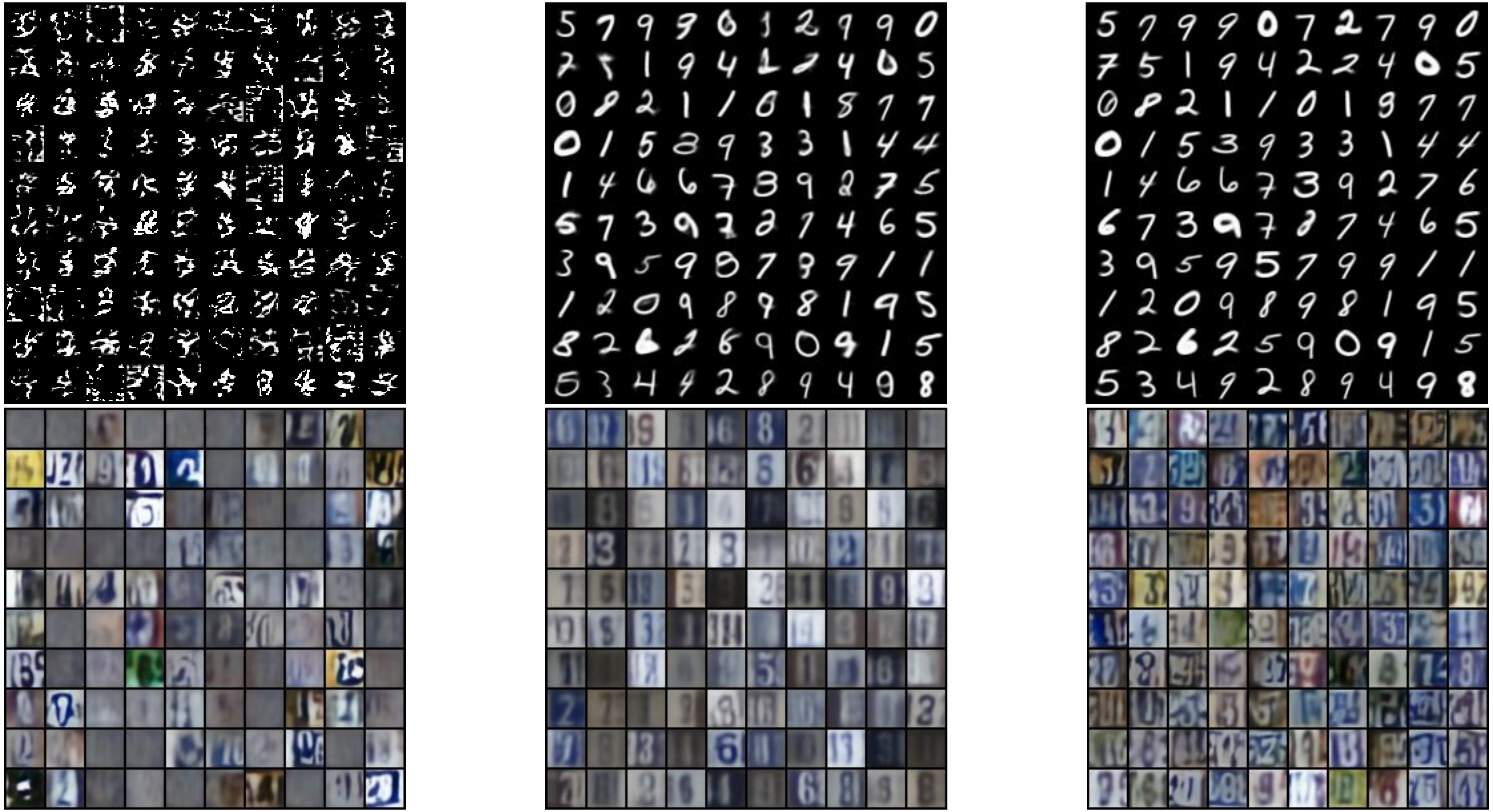
Experiment Results: MNIST & SVHN

- Data generation

DAE **VAE** **CyGen(PT)**

MNIST

SVHN



FID: 157 128 102

Experiment Results: MNIST & SVHN

- Downstream classification on the latent space:

  A hint on posterior collapse.

  - †: For BiGAN and GibbsNet, report the results in [Lamb'17] which use a deterministic architecture (failure using the same architecture).

	DAE	VAE	BiGAN†	GibbsNet†	**CyGen(PT)**
MNIST	98.0±0.1	94.5±0.3	91.0	97.7	**98.3±0.1**
SVHN	74.5±1.0	30.8±0.2	66.7	**79.6**	**75.8±0.5**

Thanks!

https://arxiv.org/abs/2106.15962

References

- [Heckerman'00] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. Journal of Machine Learning Research, 1(Oct):49-75, 2000.
- [Vincent'08] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the International Conference on Machine Learning, pages 1096-1103, 2008.
- [Grover'19] A. Grover and S. Ermon. Uncertainty autoencoders: Learning compressed representations via variational information maximization. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 2514-2524. PMLR, 2019.
- [Bengio'13] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In Advances in Neural Information Processing Systems, 2013.
- [Lamb'17] A. M. Lamb, D. Hjelm, Y. Ganin, J. P. Cohen, A. C. Courville, and Y. Bengio. GibbsNet: Iterative adversarial inference for deep graphical models. In Advances in Neural Information Processing Systems, pages 5089-5098, 2017.
- [He'16] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. Dual learning for machine translation. In Advances in Neural Information Processing Systems, pages 820-828, 2016.
- [Xia'17a] Y. Xia, J. Bian, T. Qin, N. Yu, and T.-Y. Liu. Dual inference for machine learning. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17), pages 3112-3118, 2017.
- [Xia'17b] Y. Xia, T. Qin,W. Chen, J. Bian, N. Yu, and T.-Y. Liu. Dual supervised learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3789-3798. JMLR.org, 2017.
- [Lin'19] J. Lin, Z. Chen, Y. Xia, S. Liu, T. Qin, and J. Luo. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. IEEE transactions on pattern analysis and machine intelligence, 2019.

References

- [Kim'17] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1857-1865. JMLR.org, 2017.
- [Zhu'17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 2223-2232, 2017.
- [Yi'17] Z. Yi, H. Zhang, P. Tan, and M. Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, pages 2849-2857, 2017.
- [Arnold'89] B. C. Arnold and S. J. Press. Compatible conditional distributions. Journal of the American Statistical Association, 84(405):152-156, 1989.
- [Arnold'01] B. C. Arnold, E. Castillo, J. M. Sarabia, et al. Conditionally specified distributions: an introduction. Statistical Science, 16(3):249-274, 2001.
- [Berti'14] P. Berti, E. Dreassi, and P. Rigo. Compatibility results for conditional distributions. Journal of Multivariate Analysis, 125:190-203, 2014.
- [Bengio'14] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In International Conference on Machine Learning, pages 226–234, 2014.
- [Vincent'11] P. Vincent. A connection between score matching and denoising autoencoders. Neural Computation, 23(7):1661–1674, 2011.
- [Alain'14] G. Alain and Y. Bengio. What regularized auto-encoders learn from the data-generating distribution. The Journal of Machine Learning Research, 15(1):3563–3593, 2014.

References

- [Zhang’09] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009), pages 647-655. AUAI Press, 2009.
- [Peters’14] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. Journal of Machine Learning Research, 15(1):2009–2053, 2014.
- [Shao’18] H. Shao, A. Kumar, and P. Thomas Fletcher. The Riemannian geometry of deep generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 315–323, 2018.
- [Dai’19] B. Dai and D.Wipf. Diagnosing and enhancing VAE models. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.
- [Nakagawa’21] A. Nakagawa, K. Kato, and T. Suzuki. Quantitative understanding of VAE as a non-linearly scaled isometric embedding. In Proceedings of the 38th International Conference on Machine Learning, 2021.
- [VDBerg’18] R. Van Den Berg, L. Hasenclever, J. M. Tomczak, and M. Welling. Sylvester normalizing flows for variational inference. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, pages 393–402. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.
- [Rifai’11] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the International Conference on Machine Learning, 2011.
- [Hutchinson’89] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. Communications in Statistics-Simulation and Computation, 18(3):1059-1076, 1989.