

# Boltzmann Generators

Frank Noé\* and Hao Wu

November 13, 2018

FU Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin

\* Correspondance to: frank.noe@fu-berlin.de

## Abstract

Computing equilibrium structures and statistics of condensed-matter many-body systems such as spin systems or solvated macromolecules presently relies on computationally intensive simulations such as molecular dynamics (MD) or Markov-chain Monte Carlo (MCMC) that take many small steps in configuration space. Here we combine recent progress in deep learning with principles of statistical mechanics to develop Boltzmann Generators – neural networks that learn a variable transformation such that the equilibrium distribution of a given many-body system can be “one-shot” sampled, i.e. by sampling statistically independent random variables from a simple distribution in the transformed space. As an input, the Boltzmann Generator only needs to know the potential energy function of the system, and optionally examples of known configurations. Boltzmann Generators can both compute statistics, e.g., free energy differences, of known system states or phases, and discover new ones many orders of magnitude faster than a local MD or MCMC method.

Statistical mechanics is concerned with computing the average behavior of many copies of a physical system based on models of its microscopic constituents and their interactions. For example, what is the average magnetization in an Ising model of interacting magnetic spins in an external field, or what is the probability of a protein to be folded in an atomistic molecular model as a function of the temperature. Under a wide range of conditions, the equilibrium probability of a microscopic configuration  $\mathbf{x}$  (setting of all spins, positions of all protein atoms, etc.) is proportional to  $e^{-u(\mathbf{x})}$ , e.g., the Boltzmann distribution. The dimensionless energy  $u(\mathbf{x})$  contains the potential energy of the system, the temperature and optionally other thermodynamic quantities (SI).

Except for simple model systems, we presently have no approach to directly draw statistically independent samples  $\mathbf{x}$  from the Boltzmann distribution in order to compute Boltzmann-weighted averages such as free energy differences. Therefore, currently used simulation methods are variants of Markov-Chain Monte Carlo (MCMC) or Molecular Dynamics (MD) simulations that make tiny changes to  $\mathbf{x}$  in each simulation step and will, over a long time, generate trajectories that sample from the Boltzmann distribution. These methods suffer from rare events, i.e., many simulation steps are needed to transition between different phases or metastable (long-lived) states – for example,  $10^9 - 10^{15}$  MD simulation steps are needed to fold or unfold a protein. As a result, these methods are extremely expensive and consume much of the worldwide supercomputing resources. Rare event transitions which can be traced by low-dimensional coordinates can be sped up by enhanced sampling methods [17, 6, 11], but the computational effort remains enormous.

Here we set out to develop a “Boltzmann Generator” machine that is trained on a given energy function  $u(\mathbf{x})$  and then produces statistically independent samples from  $e^{-u(\mathbf{x})}$ . At first sight, this enterprise seems hopeless for condensed-matter systems and complex polymers (e.g., Fig. 3a, Fig. 4k). In these systems, particles with strong repulsive interactions are densely packed in space, such that the number of low-energy configurations are vanishingly few compared to the number of possible ways to place particles.

Key to the solution is combining the strengths of deep machine learning [12] and statistical mechanics (Fig. 1a): We train a deep invertible neural network, to learn a coordinate transformation from  $\mathbf{x}$  to a so-called “latent” representation  $\mathbf{z}$ , in which sampling is easy and every sample can be back-transformed to a configuration  $\mathbf{x}$  with high Boltzmann probability. We can improve the ability to find relevant parts of configuration space by “learning from example”, where we feed the Boltzmann Generator not only with the potential energy  $u(\mathbf{x})$ , but also relevant samples  $\mathbf{x}$ , such as samples from the folded or unfolded state of a protein, but without knowing the probabilities of these states. Then we employ statistical mechanics which offers a rich set of tools to generate the target distribution  $e^{-u(\mathbf{x})}$  when the proposal distribution  $p_X(\mathbf{x})$  is sufficiently similar.

It is demonstrated that Boltzmann Generators can be trained to sample low-energy structures of condensed-matter systems and complex polymer structures in “one-shot”. When the Boltzmann Generator is initialized with a few structures from different metastable states, it can generate independent samples from these states and can compute the free energy difference between them without suffering from rare events. We also demonstrate that the Boltzmann Generator has a chance of generating new, previously unseen states. Exploiting this property, an “iterative discovery” procedure is constructed in which the Boltzmann Generator gradually explores the state space.

Neural networks that can draw statistically independent samples from a desired distribution are called directed generative networks [5, 8]. Such generative networks have been demonstrated to draw photo-realistic images [7], to produce deceptively realistic speech audio [15], and even to sample formulae of chemical compounds with certain physico-chemical properties [4]. In these domains, the exact target distribution is not known and the network is “trained by example” using large databases of images, audio or molecules. Here we are in the inverse situation, as we can compute the Boltzmann weight of each generated sample  $\mathbf{x}$ , but we do not have samples from the Boltzmann distribution *a priori*. The idea of Boltzmann Generators is as follows:

1. We learn a neural network transformation  $F_{zx}$  such that when sampling from a simple distribution in  $\mathbf{z}$ , such as a Gaussian normal distribution,  $F_{zx}(\mathbf{z})$  will provide a configuration  $\mathbf{x}$  which has a high Boltzmann weight, i.e. is coming from a distribution  $p_X(\mathbf{x})$  that is similar to the target Boltzmann distribution (Fig. 1).
2. To compute Boltzmann-weighted averages, we reweight the generated distribution  $p_X(\mathbf{x})$  to the Boltzmann distribution  $e^{-u(\mathbf{x})}$ . This can be achieved with various algorithms; here the simplest one is used: assign the statistical weight  $w(\mathbf{x}) = e^{-u(\mathbf{x})}/p_X(\mathbf{x})$  to every sample  $\mathbf{x}$  and then compute any desired averages using this weight.

For both, training and reweighting, it is important that we can compute the probability weight  $p_X(\mathbf{x})$  of every sample  $\mathbf{x}$ . This can be achieved when  $F_{zx}$  is an invertible transformation, for which we can compute  $p_X(\mathbf{x})$  from the known  $p_Z(\mathbf{z})$  (Fig. 1, SI). Invertible neural network transformations are similar to flows of a fluid that transform the probability density around from configuration space to latent space. Here we consider invertible neural network blocks that are volume-preserving (as in

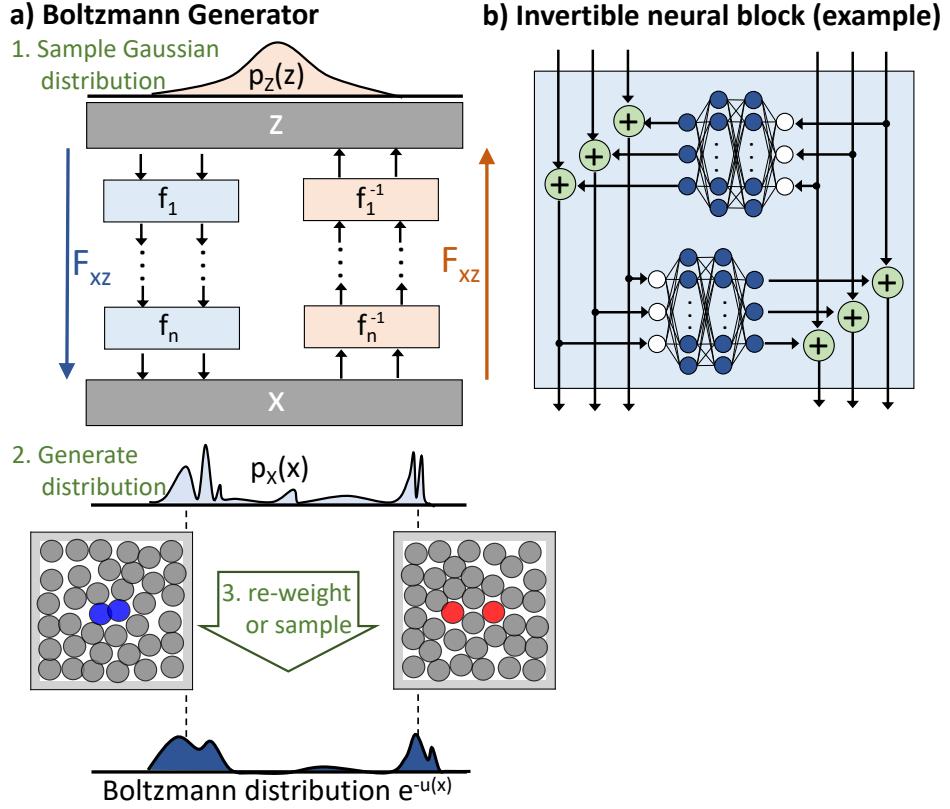


Figure 1: **Boltzmann Generators.** **a)** A Boltzmann Generator is trained by minimizing the difference between its generated distribution and the desired Boltzmann distribution. It is used by transforming samples from a simple (e.g., Gaussian) distribution to generated configurations. To compute thermodynamics, such as configurational free energies, the samples must be reweighted to the Boltzmann distribution. **b)** The Boltzmann Generator is composed of invertible neural network blocks. Here, a volume-preserving block is shown as an example.

incompressible fluids) [2], and non-volume preserving (as in compressible fluids) [10] (Suppl. Fig. 5b-e). Invertibility is achieved by special neural network architectures (Fig. 1b, Suppl. Fig. 5c,e; see SI for details). Invertible blocks are stacked to form a deep invertible neural network (Fig. 1a, Suppl. Fig. 1f).

Boltzmann Generators are trained by combining two modes: *training by energy* and *training by example*. In training by energy, it is exploited that the target distribution is known up to a constant,  $e^{-u(\mathbf{x})}$ , and the Boltzmann Generator learns to sample from a distribution similar to this target by minimizing the following loss function (SI):

$$J_{KL} = \mathbb{E}_{\mathbf{z} \sim q_Z(\mathbf{z})} [u_X(F_{zx}(\mathbf{z})) - S_{zx}(\mathbf{z})] \quad (1)$$

Here,  $u_X(F_{zx}(\mathbf{z}))$  takes the Gaussian random variable  $\mathbf{z}$ , transforms it to a configuration  $\mathbf{x}$  and computes its energy.  $S_{zx}$  quantifies how much the network scales the configuration space volume at  $\mathbf{z}$  and can be computed from the scaling transformations of the network (SI). Training by energy can be extended to train for the Boltzmann distribution at a range of temperatures, by exploiting that the temperature, which changes the configuration space distribution in a complex way, simply enters as a scaling factor in the width of the Gaussian  $q_Z(\mathbf{z})$  (SI).

In Machine Learning areas where the target distribution is unknown, one typically employs training

by example, e.g. using the maximum likelihood principle [2]. Training by example is also useful here, in order to guide the Boltzmann Generator to sample regions of configuration space that are relevant. Examples for configurations  $\mathbf{x}$  may come from some initial simulation data, e.g. short MD simulations in one or a few metastable states, that are not (yet) Boltzmann distributed. We also use this training mode for iterative discovery: Samples  $\mathbf{x}$  that have been generated previously and that are in new metastable states can be used as examples in a new round of training, in order to force the Boltzmann Generator to keep generating from known metastable states. Training by example works by sampling configurations  $\mathbf{x}$  from the input data, transforming each of them to the latent space by  $\mathbf{z} = F_{xz}(\mathbf{x})$ , and minimizing the likelihood loss:

$$J_{ML} = \mathbb{E}_{\mathbf{x}} [u_Z(F_{xz}(\mathbf{x})) - S_{xz}(\mathbf{x})]. \quad (2)$$

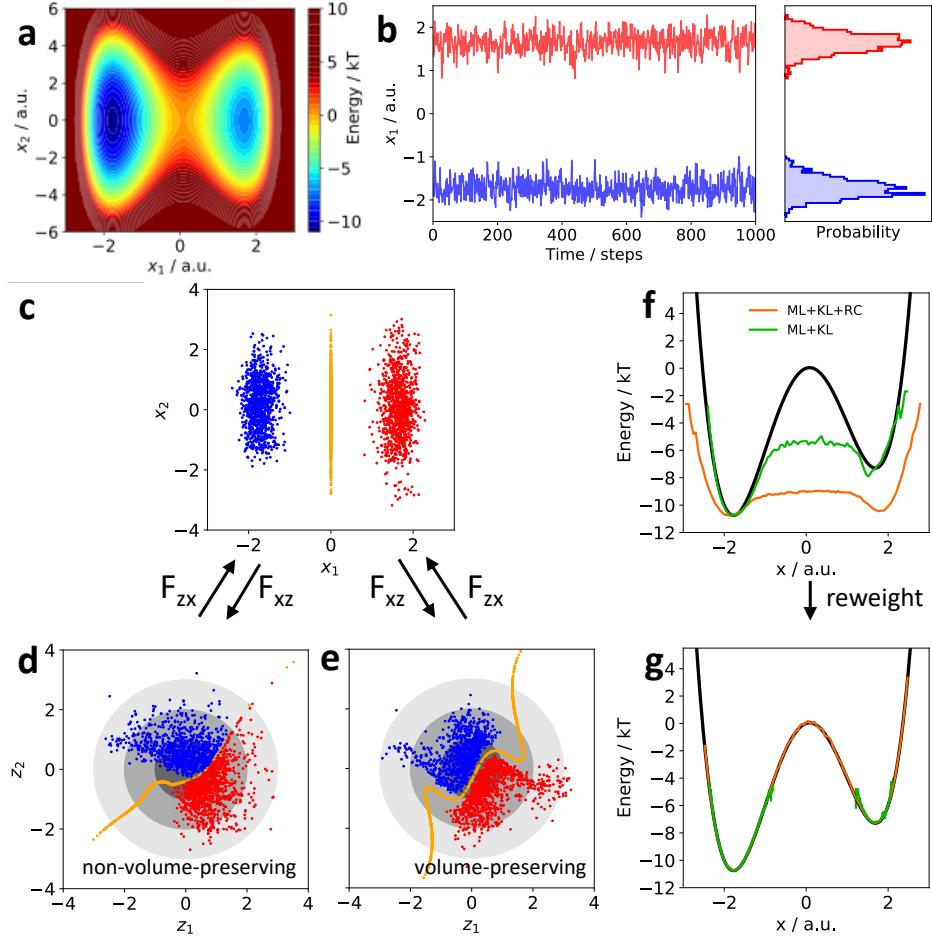
In both (1-2), the first term is minimizing the mean potential energy, i.e. the enthalpy of the system, and the second term is maximizing the entropy of the transformation, which avoids that the network only learns to sample the minimum energy configuration. Still, minimizing only  $J_{KL}$  tends to focus sampling on the most stable metastable state (Suppl. Fig. 6,7). We found it beneficial to combine both losses, in particular it is useful to start the training by minimizing  $J_{ML}$  using a small set of simulation data in order to initialize  $T_{zx}$  to point to the relevant part of state space.

By combining training by energy and training by example, we can sample configurations that have high probabilities and low free energies. However, sometimes we want to generate certain states with a low probability, for example the transition states along a certain reaction coordinate (RC) along which we want to compute the free energy profile. Standard sampling methods, such as MD and MCMC, can be combined with Umbrella Sampling [17], Metadynamics [11] or Flooding [11, 6] in order to bias the sampled distribution to be more uniform along a chosen RC. For the same purpose, we introduce an RC loss that can optionally be used to enhance the sampling of a Boltzmann Generator along a chosen RC (SI).

We first illustrate Boltzmann Generators using a two-dimensional potential that has two stable states separated by a high energy barrier in  $x_1$ -direction, while it is a harmonic oscillator in  $x_2$  (Fig. 2a). MD simulations will oscillate close to one of the stable state for a long time before a rare transition event occurs (Fig. 2b). Hence, the distribution in configuration space  $(x, y)$  is split into two modes (Fig. 2c, transition state ensemble is shown in yellow for clarity but not used for training). We are training Boltzmann Generators using the two short and disconnected simulations shown in Fig. 2b as example. Fig. 2d,e show the latent space learned by non-volume-preserving and a volume-preserving transformation, respectively. In both cases, the probability densities of the two states and the transition state are “repacked” so as to form a compact density close to the center.

We use the Boltzmann generator by sampling from its latent space according to the Gaussian distribution. After transforming these variables via  $T_{zx}$ , this produces uncorrelated samples from both stable states without any sampling problem. A variety of training methods succeed in sampling across the barrier such that the rare event nature of the system is eliminated (Suppl. Figs. 6,7). Combining a Boltzmann Generator trained by minimizing (1-2) with simple reweighting reproduces the precise free energy differences of the two metastable states (Fig. 2g green). By additionally training with the RC loss to promote sampling along  $x_1$ , the low-probability transition states are sampled (Fig. 2f orange), and the full free energy profile along  $x_1$  can be reconstructed with high precision (Fig. 2f,g orange).

For the double-well system, the unbiased MD simulation needs on average  $4 \cdot 10^6$  MD steps for a return trip between the two states (SI). The Boltzmann Generator replaces this by statistically independent



**Figure 2: Illustration of Boltzmann Generators for two-dimensional bistable system.** **a)** Two-dimensional potential,  $x_1$  is the slow coordinate. **b)** Two short simulation trajectories that stay in their metastable states without crossing. **c)** Distribution of trajectories of b) in configuration space ( $x_1, x_2$ ). Transition state ensemble is shown (orange) but not used for training. **d,e)** Latent-space distribution of trajectories of b) when mapped through trained  $F_{xz}$  using non-volume preserving (d) and volume-preserving (e) transformations. **f)** Free energy corresponding to distribution sampled by Boltzmann Generators trained mainly with Eq. (1) (green) and using reaction coordinate training (orange). **g)** Free energy estimates obtained after reweighting, colors as in (f).

samples, corresponding to a speedup by up to six orders of magnitude.

As a second example, we demonstrate that Boltzmann Generators can sample high-probability structures and efficiently compute the thermodynamics in crowded condensed matter systems. We simulated a dense system of two-dimensional particles confined to a box as suggested in [14] (Fig. 3a). Immersed in the fluid is a bistable particle dimer whose open and closed states are separated by a high barrier (Fig. 3a-c). Directly opening or closing the dimer is not possible due to the high density of the system and requires a concerted rearrangement of the solvent particles. Particles repel each other with the 12<sup>th</sup> power of their inverse distance when being close. As a result, the fraction of low-energy configurations is vanishingly small, and designing a sampling method by hand that would simultaneously place all 38 particles and achieve low energies appears unfeasible.

We train a Boltzmann Generator to sample low-energy configurations in “one-shot” and use it in order to compute the free energy profiles of opening / closing the dimer. The training is initialized with examples from separate, disconnected simulations of the open and closed states, but in later stages, mostly training by energy (1) is used. A restraint keeps the bistable particle dimer centered and aligned in the simulation box, therefore the  $x$ -position of each dimer particle indicates if we are in the open or closed state (Fig. 3d). The trained Boltzmann Generator has learned a transformation of the complex configuration space density to a compact, 76-dimensional ball in latent space (Fig. 3e). Direct sampling of from 76-dimensional Gaussian in latent space and transformation via  $T_{zx}$  generates configurations where all particles are placed without significant clashes, and potential energies that overlap with the energy distribution of the unbiased MD trajectories (Fig. 3f-h). Also, realistic transition states that have not been included in any training data are sampled (Fig. 3g).

We estimate that the MD simulation needs at least  $10^{12}$  steps to spontaneously see a transition from closed to open state and back (SI). The Boltzmann generator samples these states with a return time of about 4 steps, corresponding to more than 11 orders of magnitude speedup.

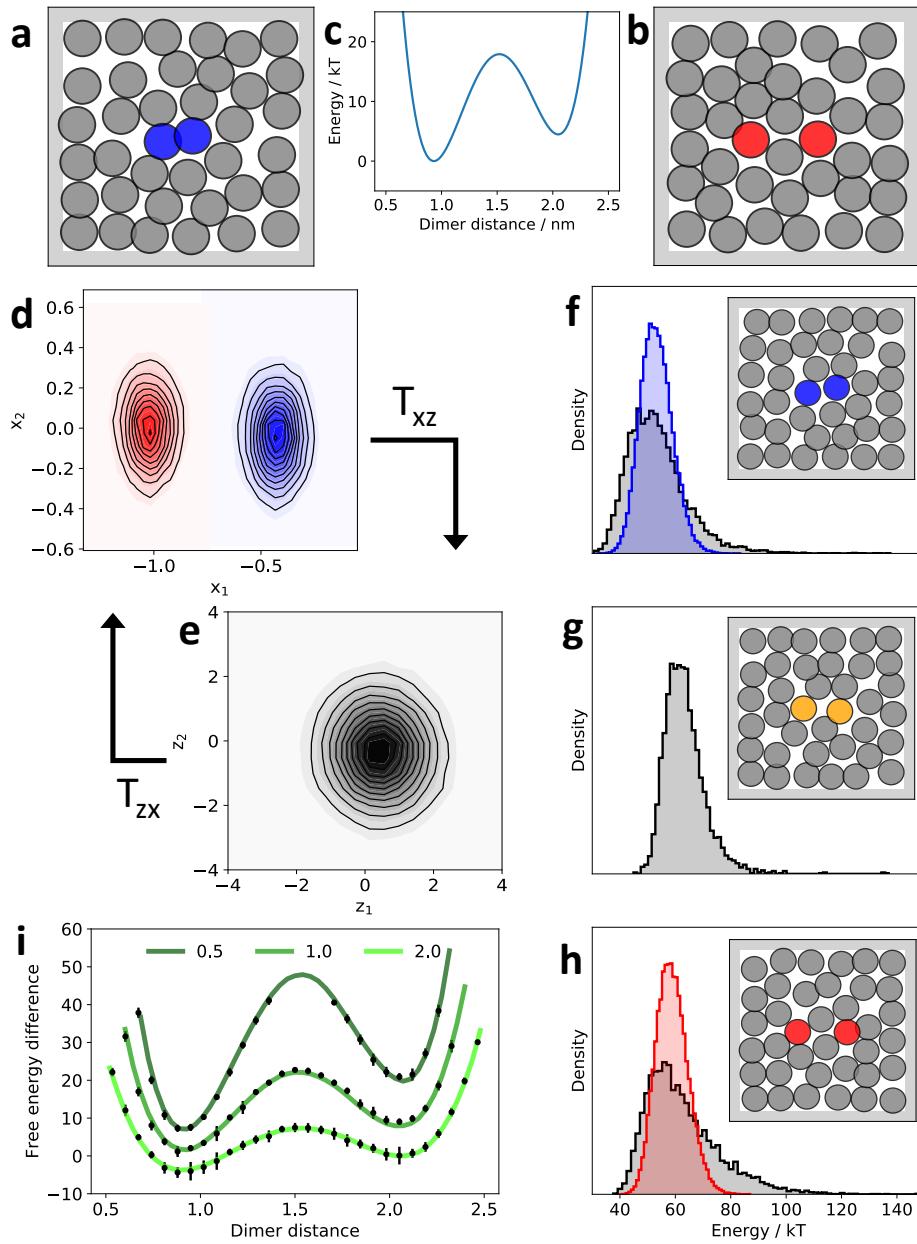
To demonstrate that thermodynamic quantities can be computed with Boltzmann Generators, we perform the training by energy (1) simultaneously to a range of temperatures between one fourth and four times the reference temperature (SI). Then, we sample the Boltzmann Generator for a range of temperatures and use simple reweighting to compute the free energies along the dimer distances. As shown in Fig. 3i, these temperature-dependent free energies agree precisely with extensive umbrella sampling simulations that employ bias potentials along the dimer distance [17].

Finally, we demonstrate that Boltzmann Generators can sample complex polymer structures in “one shot” that belong to known or new metastable states. Cyclical polymers are especially challenging, because the main degrees of freedom are torsion angles, but for each change of a torsion, other torsions must be changed concurrently so as to maintain ring closure and all bond angle constraints. Sophisticated Monte Carlo moves have been designed for this purpose [3], but they generally do not yield MCMC procedures that sample the Boltzmann distribution.

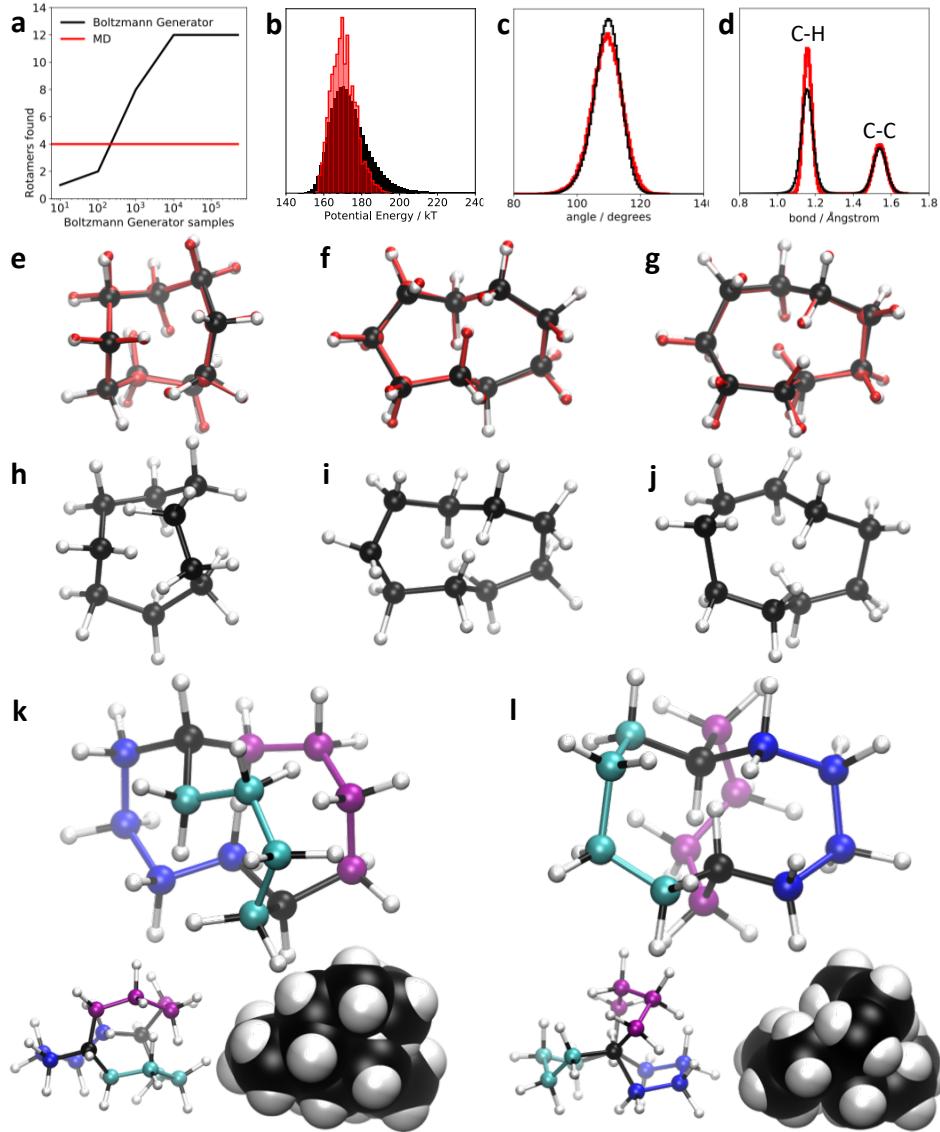
Here we use cyclical hydrocarbons as example. Each hydrocarbon torsion angle has three rotamers (around  $-60^\circ$ ,  $60^\circ$ ,  $180^\circ$ ). The cycle constraints stabilize some otherwise unstable conformations but generally reduce the total number of accessible conformations. For cyclononane ( $C_9H_{18}$ ), we used a combination of training by energy and training by example, the latter initialized with a short replica-exchange MD simulation in which 4 distinct rotamers conformations have been sampled. We then use a Boltzmann Generator for iterative discovery: In each iteration, the Boltzmann generator samples structures from known conformation, and also a small fraction of structures from new conformations. We resample this result to produce a uniform number of samples from each conformation and use these samples for training by example in the next round (SI). A so-trained Boltzmann Generator quickly produces structures not included in the initial MD data (Fig. 4a, e-j). Potential energies over generated structures have a high overlap with the potential energy sampled in the MD simulation (Fig. 4b). Note that all atoms – including hydrogens – are generated in one shot. Nonetheless, the bond lengths and bond angles follow their equilibrium distribution closely (Fig. 4c,d).

Finally, one-shot samples were generated for bicyclo[4.4.4]tetradecane  $C_{14}H_{26}$ , a highly constrained and densely packed hydrocarbon with two interconnected ring systems and 120 dimensions (Fig. 4k,l and insets).

Boltzmann Generators are, as yet, the first approach that can sample the Boltzmann distribution and generate structures of condensed matter systems and complex polymers directly, i.e. by avoiding to



**Figure 3: Particle System with bistable dimer.** **a,b)** Closed (blue) and open (red) configurations from MD simulations (input data). **c)** Bistable dimer potential. **d)** Distribution of input data on  $x_1, x_2$ . **e)** Distribution of input data on  $z_1, z_2$  after training. **f, g, h)** Potential energy distribution from MD (colored) and Boltzmann generator (grey) for closed (f), open (h) and transition configurations (g). Insets show one-shot samples from Boltzmann generator. **i)** Free energy differences as a function of dimer distance and temperature sampled with Boltzmann generators (one-shot generation and reweighting, bullets with error bars indicating one standard deviation) and umbrella sampling (green lines).



**Figure 4: One-shot sampling of cyclical molecule structures:** Cyclononane  $C_9H_{18}$  (a-j) and bicyclo[4.4.4]tetradecane  $C_{14}H_{26}$  (k-l). **a)** Number of distinct rotamers samples with Boltzmann Generator that is initialized with MD data containing 12 rotamers. **b)** Potential energy distribution. **c-d)** Generated bond length and angle distribution compared to MD data. **e-g)** One-shot Boltzmann-generated structures (black) and the most similar MD structures from replica-exchange MD (red). **h-j)** One-shot Boltzmann-generated structures that are not contained in the MD simulations. **k-l)** One-shot Boltzmann-generated structures of bicyclo[4.4.4]tetradecane  $C_{14}H_{26}$  – side-view is shown in large, top view below as ball+stick and space-filling representation. The three chains are colored for better visibility, the connecting carbons are kept dark.

make small MD or MCMC steps. We have demonstrated this for systems with around 100 dimensions. Although we expect the methodology to improve rapidly, we believe that for very high-dimensional systems, such as solvated atomistic protein models with 100,000’s of dimensions, the best strategy is to employ Boltzmann Generators for Monte Carlo cluster moves. By grouping the system into clusters of atoms, and can re-sample these clusters according to their marginal Boltzmann distribution, i.e. using the internal potential energies and the interaction energies between the cluster and the remaining system. With such a strategy, Boltzmann Generators can be naturally combined with existing local sampling methods.

The present work shows that Boltzmann Generators can sample the Boltzmann distribution of complex systems in one shot and may offer a way out of the sampling problem in condensed matter systems. The limitation of the current work is that the transformation that achieves this needs to be trained using the system-specific energy. In order to make the approach general, it needs to be transferrable across systems, and a promising route is to employ transferrable featurization methods developed in the context of Machine Learning for Quantum Mechanics [1, 16].

**Acknowledgements** FN acknowledges funding from European Commission (ERC CoG 772230 “ScaleCell”), Deutsche Forschungsgemeinschaft (CRC1114/A04), and the MATH<sup>+</sup> research cluster (AA1x8, EF1x2).

## References

- [1] J. Behler and M. Parrinello. “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces.” In: *Phys. Rev. Lett.* 98 (2007), p. 146401.
- [2] Laurent Dinh, David Krueger, and Yoshua Bengio. “NICE: Nonlinear independent components estimation.” In: *ICLR* (2015), arXiv:1410.8516.
- [3] N. Go and H. A. Scheraga. “Ring closure and local conformational deformations of chain molecules.” In: *Macromolecules* 3 (1970), pp. 178–187.
- [4] Rafael Gómez-Bombarelli et al. “Automatic chemical design using a data-driven continuous representation of molecules.” In: *ACS Cent. Sci.* 4 (2018), pp. 268–276.
- [5] I. Goodfellow et al. “Generative Adversarial Networks.” In: *arXiv:1406.2661* (2014).
- [6] H. Grubmüller. “Predicting slow structural transitions in macromolecular systems: conformational flooding.” In: *Phys. Rev. E* 52 (1995), p. 2893.
- [7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. “Progressive Growing of GANs for Improved Quality, Stability, and Variation.” In: *ICLR* (2018).
- [8] D. P. Kingma and M. Welling. “Auto-Encoding Variational Bayes.” In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, arXiv:1312.6114. 2014.
- [9] Harold W. Kuhn. “The Hungarian Method for the assignment problem.” In: *Nav. Res. Logist. Quart.* 2 (1955), pp. 83–97.
- [10] S. Bengio L. Dinh J. Sohl-Dickstein. “Density estimation using Real NVP.” In: *arXiv:1605.08803* (2016).
- [11] A. Laio and M. Parrinello. “Escaping free energy minima.” In: *Proc. Natl. Acad. Sci. USA* 99 (2002), pp. 12562–12566.

- [12] Y. LeCun, Y. Bengio, and G. Hinton. “Deep learning.” In: *Nature* 521 (2015), pp. 436–444.
- [13] Alexei M. Nikitin, Yury V. Milchevskiy, and Alexander P. Lyubartsev. “A new AMBER-compatible force field parameter set for alkanes.” In: *J. Mol. Model.* 20 (2014), p. 2143.
- [14] Jerome P. Nilmeier, Gavin E. Crooks, David D. L. Minh, and John D. Chodera. “Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation.” In: *Proc. Natl. Acad. Sci. USA* 108.E1009-E1018 (2011).
- [15] Aaron van den Oord et al. “Parallel WaveNet: Fast High-Fidelity Speech Synthesis.” In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, 2018, pp. 3918–3926.
- [16] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld. “Fast and accurate modeling of molecular atomization energies with machine learning.” In: *Phys. Rev. Lett.* 108 (2012), p. 058301.
- [17] G. M. Torrie and J. P. Valleau. “Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling.” In: *J. Comp. Phys.* 23 (1977), pp. 187–199.

# Supplementary Material

## A. Invertible networks

We employ invertible networks in order to learn the transformation between the Gaussian random variables  $\mathbf{z}$  and the Boltzmann-distributed random variables  $\mathbf{x}$ :

$$\begin{aligned}\mathbf{z} &= F_{xz}(\mathbf{x}; \boldsymbol{\theta}) \\ \mathbf{x} &= F_{zx}(\mathbf{z}; \boldsymbol{\theta}).\end{aligned}$$

Hence  $T_{xz} = T_{zx}^{-1}$ . Note that the set of parameters  $\boldsymbol{\theta}$  defining these transformations are identical (shared) between the forward and backward transformations. Each transformation has a Jacobian matrix with the pairwise first derivatives of outputs with respect to inputs:

$$\begin{aligned}\mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta}) &= \left[ \frac{\partial F_{zx}(\mathbf{z}; \boldsymbol{\theta})}{\partial z_1}, \dots, \frac{\partial F_{zx}(\mathbf{z}; \boldsymbol{\theta})}{\partial z_n} \right] \\ \mathbf{J}_{xz}(\mathbf{x}; \boldsymbol{\theta}) &= \left[ \frac{dF_{xz}(\mathbf{x}; \boldsymbol{\theta})}{dx_1}, \dots, \frac{dF_{xz}(\mathbf{x}; \boldsymbol{\theta})}{dx_n} \right]\end{aligned}$$

The absolute value of the Jacobian's determinant,  $|\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})|$ , measures how much a volume element at  $\mathbf{z}$  is scaled by the transformation. Forward and reverse transformation are related by  $|\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})| = |\det \mathbf{J}_{xz}(\mathbf{x})|^{-1}$ , and respectively for  $\mathbf{x}$  and  $\mathbf{z}$  exchanged. As we frequently deal with logarithms of Jacobian determinants, we introduce the abbreviations:

$$\begin{aligned}S_{xz}(\mathbf{x}) &= \log |\det \mathbf{J}_{xz}(\mathbf{x})| \\ S_{zx}(\mathbf{z}) &= \log |\det \mathbf{J}_{zx}(\mathbf{z})|.\end{aligned}$$

Our main motivation to use invertible transformations is that they allow us to transform random variables as follows:

$$\begin{aligned}p_X(\mathbf{x}) &= p_Z(\mathbf{z}) |\det \mathbf{J}_{zx}(\mathbf{z})|^{-1} \\ &= p_Z(T_{xz}(\mathbf{x})) |\det \mathbf{J}_{xz}(\mathbf{x})| \tag{3}\end{aligned}$$

$$\begin{aligned}p_Z(\mathbf{z}) &= p_X(\mathbf{x}) |\det \mathbf{J}_{xz}(\mathbf{x})|^{-1} \\ &= p_X(T_{zx}(\mathbf{z})) |\det \mathbf{J}_{zx}(\mathbf{z})| \tag{4}\end{aligned}$$

Here we employ the invertible network structures NICE [2] and RealNVP [10]. The main idea is to split the variables into two channels,  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$  and  $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ , do only trivially invertible operations on each channel, such as multiplication and addition, and use trainable, nonlinear neural network transformations between the channels to compute the value of these multiplication and addition transformations (Suppl. Fig. 5b-e).

Table 1 summarizes the transformations employed here, their inverses and Jacobian determinant values. A single transformation of  $(\mathbf{z}_1, \mathbf{z}_2) = f_{xz}(\mathbf{x}_1, \mathbf{x}_2)$ , where  $T_{xz}$  is implemented via NICE or RealNVP transforms only the second channel and leaves the first channel unchanged. In order to allow all variables to be transformed, we swap channels in the next transformation (Suppl. Fig. 5f), and define

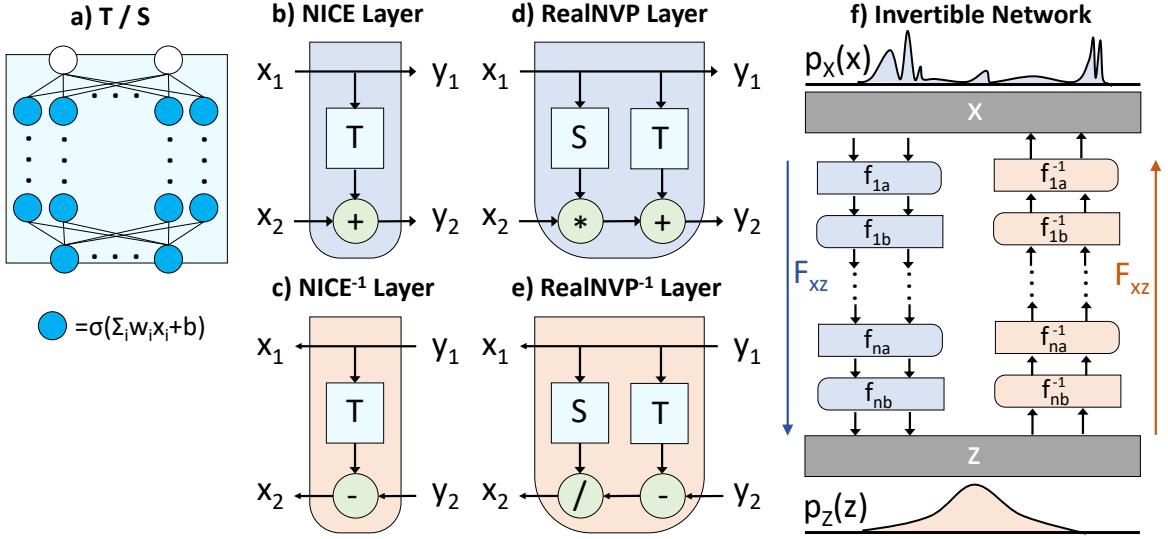


Figure 5: **Boltzmann Generator network architecture.** a) Nonlinear transformations  $T$  and  $S$  are built with multilayer neural networks. b,c) Volume-preserving NICE layer and its inverse. d,e) Non-volume-preserving RealNVP layer and its inverse. f) Stacking any sequence of these layers with channels exchanges produces the full Boltzmann Generator, and invertible network.

Layer	$f_{xz}$	$ \det \mathbf{J}_{xz} $	$f_{zx}$	$ \det \mathbf{J}_{zx} $
NICE	$\mathbf{z}_1 = \mathbf{x}_1$ $\mathbf{z}_2 = \mathbf{x}_2 + T(\mathbf{x}_1; \theta)$	1	$\mathbf{x}_1 = \mathbf{z}_1$ $\mathbf{x}_2 = \mathbf{z}_2 - T(\mathbf{y}_1; \theta)$	1
Scaling, Exp	$\mathbf{z} = e^{\mathbf{k}} \circ \mathbf{x}$	$e^{\sum_i k_i}$	$\mathbf{x} = e^{-\mathbf{k}} \circ \mathbf{z}$	$e^{-\sum_i k_i}$
RealNVP	$\mathbf{z}_1 = \mathbf{x}_1$ $\mathbf{z}_2 = \mathbf{x}_2 \odot \exp(S(\mathbf{x}_1; \theta)) + T(\mathbf{x}_1; \theta)$	$e^{\sum_i S_i(\mathbf{x}_1; \theta)}$	$\mathbf{x}_1 = \mathbf{z}_1$ $\mathbf{x}_2 = (\mathbf{z}_2 - T(\mathbf{x}_1; \theta)) \odot \exp(-S(\mathbf{z}_1; \theta))$	$e^{-\sum_i S_i(\mathbf{z}_1; \theta)}$

Table 1: Invertible network components.

a NICE or RealNVP block as:

$$(\mathbf{y}_1, \mathbf{y}_2) = f_{xy}(\mathbf{x}_1, \mathbf{x}_2)$$

$$(\mathbf{z}_1, \mathbf{z}_2) = f_{yz}(\mathbf{y}_2, \mathbf{y}_1)$$

Boltzmann Generators are build by putting the forward and the inverse of such blocks in parallel (Fig. 1f). The forward and the inverse transformation in each layer share the same nonlinear transformation ( $T$  or  $S$ ), and therefore the same parameters.

The NICE transformation is volume-preserving. As such, it also preserves the entropy  $H_X = \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$ . In order to be able to model probability distributions with arbitrary entropy, we need to insert at least one scaling layer into a Boltzmann Generator that otherwise only contains NICE layers.

## B. Training and using Boltzmann Generators

The Boltzmann Generator is trained by minimizing a loss functional that has the following form:

$$J = w_{ML} J_{ML} + w_{KL} J_{KL} + w_{RC} J_{RC}.$$

where the terms represent maximum-likelihood (ML, “training by example”), Kullback-Leiber (KL, “training by energy”), and reaction-coordinate (RC) optimization and the  $w$ ’s control their weights. Below we will derive these terms in detail.

We call the “exact” distributions  $\mu$  and the generated distributions  $q$ . In particular,  $\mu_Z(\mathbf{z})$  is the Gaussian prior distribution injected into the latent space and  $q_X(\mathbf{x})$  is the distribution that results from the network transformation  $F_{zx}$ . Likewise,  $\mu_X(\mathbf{x}) \propto \exp(-u(\mathbf{x}))$  is the Boltzmann distribution in configuration space and  $q_Z(\mathbf{z})$  is the distribution that results from the network transformation  $F_{xz}$ :

$$\begin{array}{ccc} \mu_Z(\mathbf{z}) & \xrightarrow{F_{zx}} & q_X(\mathbf{x}) \\ \mu_X(\mathbf{x}) & \xrightarrow{F_{xz}} & q_Z(\mathbf{z}) \end{array}$$

**Boltzmann distribution:** A special case is to use Boltzmann Generators to sample from the Boltzmann distribution of the canonical ensemble. This distribution has the form:

$$\mu_X(\mathbf{x}) = Z_X^{-1} e^{-\beta U(\mathbf{x})} \quad (5)$$

where  $\beta^{-1} = k_B T$  with Boltzmann constant  $k_B$  and temperature  $T$ . When we only have one temperature, we can simply subsume the constant into a reduced energy

$$u(\mathbf{x}) = \frac{U(\mathbf{x})}{k_B T}$$

In order to evaluate a set of temperatures  $(T^1, \dots, T^K)$ , we can define a reference temperature  $T^0$  and the respective reduced energy  $u^0(\mathbf{x}) = U(\mathbf{x})/k_B T^0$  and we then obtain the reduced energies simply by scaling:

$$u^k(\mathbf{x}) = \frac{T^0}{T^k} u^0(\mathbf{x}) = \frac{u^0(\mathbf{x})}{\tau_k}$$

where  $\tau_k$  is the relative temperature.

**Prior distribution:** We sample the input in  $\mathbf{z}$  from the isotropic Gaussian distribution:

$$\mu_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} e^{-\frac{1}{2} \|\mathbf{z}\|^2 / \sigma_k^2}, \quad (6)$$

with normalization constant  $Z_Z$ . The prior energy, i.e. the energy whose Boltzmann distribution is the prior distribution, is given by:

$$\begin{aligned} u_Z^k(\mathbf{z}) &= -\log \mu_Z^k(\mathbf{z}) \\ &= \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const.} \end{aligned} \quad (7)$$

Thus the variance takes the same role as the relative temperature. We define (arbitrarily) to set the

variance equal 1 at the standard temperature, and obtain:

$$\sigma_k^2 = \tau_k.$$

**Latent KL divergence** The KL divergence between two distributions  $q$  and  $p$  is given by

$$\begin{aligned} \text{KL}(q \parallel p) &= \int q(\mathbf{x}) [\log q(\mathbf{x}) - \log p(\mathbf{x})] d\mathbf{x}, \\ &= -H_q - \int q(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}, \end{aligned}$$

where  $H_q$  is the entropy of the distribution  $q$ .

Here we use KL divergences to minimize the difference between the probability densities predicted by the Boltzmann generator and the respective reference distribution. Using the variable transformations (3-4) and the Boltzmann distribution (5), we can express the KL divergence in latent space as:

$$\begin{aligned} \text{KL}_{\boldsymbol{\theta}} [\mu_Z \parallel q_Z] &= -H_Z - \int \mu_Z(\mathbf{z}) \log q_Z(\mathbf{z}; \boldsymbol{\theta}) d\mathbf{z}, \\ &= -H_Z - \int \mu_Z(\mathbf{z}) [\log \mu_X(F_{zx}(\mathbf{z}; \boldsymbol{\theta})) + \log |\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})|] d\mathbf{z}, \\ &= -H_Z + \log Z_X + \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [u(F_{zx}(\mathbf{z}; \boldsymbol{\theta})) - \log |\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})|] \end{aligned}$$

This is equivalent to the KL divergence expressed in configuration space:

$$\begin{aligned} \text{KL}_{\boldsymbol{\theta}} [q_X \parallel \mu_X] &= \int q_X(\mathbf{x}; \boldsymbol{\theta}) [\log q_X(\mathbf{x}; \boldsymbol{\theta}) + \log Z_X + u(\mathbf{x})] d\mathbf{x} \\ &= \int \mu_Z(\mathbf{z}) [\log \mu_Z(\mathbf{z}) + \log |\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})| + \log Z_X + u(F_{zx}(\mathbf{z}; \boldsymbol{\theta}))] d\mathbf{z} \\ &= -H_Z + \log Z_X + \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [u(F_{zx}(\mathbf{z}; \boldsymbol{\theta})) - \log |\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})|] \end{aligned}$$

Here,  $\boldsymbol{\theta}$  are the trainable neural network parameters. Since  $H_Z$  and  $Z_X$  are constants in  $\boldsymbol{\theta}$ , the KL loss is given by:

$$J_{KL} = \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [u(F_{zx}(\mathbf{z}; \boldsymbol{\theta})) - \log |\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})|]. \quad (8)$$

Practically, each training batch samples  $\mathbf{z} \sim q_Z(\mathbf{z})$  from a normal distribution, transforms them via  $T_{zx}$ , and evaluates Eq. (8).

We can extend (8) to simultaneously train at multiple temperatures, obtaining:

$$J_{KL}^{T^1, \dots, T^K} = \sum_{k=1}^K \mathbb{E}_{\mathbf{z} \sim \mu_Z^k(\mathbf{z})} [u^k(F_{zx}(\mathbf{z}; \boldsymbol{\theta})) - \log |\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})|].$$

The KL divergence  $\text{KL}_{\boldsymbol{\theta}} [\mu_Z \parallel q_Z]$  is also maximized in *probability density distillation* used in the training of recent audio generation networks [15]. Here, the reference distribution is defined by a teacher network that is used to help training a student network. However, the resulting expressions are different because the target distribution is not defined by a physical energy as here.

**Reweighting and interpretation of latent KL as reweighting loss** The most direct way to compute quantitative statistics using Boltzmann generators is to employ reweighting of probability densities. In this framework, we assign to each generated configuration  $\mathbf{x}$  the statistical weight:

$$\begin{aligned} w_X(\mathbf{x}) &= \frac{\mu_X(\mathbf{x})}{q_X(\mathbf{x})} = \frac{q_Z(\mathbf{z})}{\mu_Z(\mathbf{z})}. \\ &\propto e^{-u_X(T_{zx}(\mathbf{z})) + u_Z(\mathbf{z}) + \log |\det(\mathbf{J}_{zx}(\mathbf{z}))|} \end{aligned} \quad (9)$$

where the equivalence on the right hand side results from (3-4). Using these weights, expectation values can be computed as

$$\mathbb{E}[O] \approx \frac{\sum_{i=1}^N w_X(\mathbf{x}) O(\mathbf{x})}{\sum_{i=1}^N w_X(\mathbf{x})}.$$

With the reweighting (9), the KL loss (8) has an interesting thermodynamic interpretation. The minimization of the latent KL divergence can be rewritten in terms of these weights:

$$\begin{aligned} \min \text{KL}_\theta [\mu_Z \| q_Z] &= \min \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [\log \mu_Z(\mathbf{z}) - \log q_Z(\mathbf{z}; \theta)] \\ &= \max \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [\log w_X(\mathbf{x} | \mathbf{z})]. \end{aligned}$$

Thus, the minimization of the latent KL divergence is equivalent to maximizing the expected log-weights of points, or equivalently the product of all weights, in a reweighting procedure. Indeed the maximum weights are achieved when the proposal distribution is identical to the Boltzmann distribution, resulting in  $w_X(\mathbf{x}) \equiv 1$ .

**Interpretation of latent KL as free energy** For invertible transformation  $F_{xz}$ , we additionally use the following relationship of the entropies of the two distributions:

$$\begin{aligned} H_X &= - \int_{\mathbf{x}} q_X(\mathbf{x}) \log q_X(\mathbf{x}) d\mathbf{x} \\ &= - \int_{\mathbf{z}} q_X(F_{zx}(\mathbf{z})) \log q_X(F_{zx}(\mathbf{z})) |\det \mathbf{J}_{zx}(\mathbf{z})| d\mathbf{z} \\ &= - \int_{\mathbf{z}} \mu_Z(\mathbf{z}) \log q_X(F_{zx}(\mathbf{z})) d\mathbf{z} \\ &= - \int_{\mathbf{z}} \mu_Z(\mathbf{z}) \log (\mu_Z(\mathbf{z}) |\det \mathbf{J}_{zx}(\mathbf{z})|^{-1}) d\mathbf{z} \\ &= - \int_{\mathbf{z}} \mu_Z(\mathbf{z}) \log \mu_Z(\mathbf{z}) d\mathbf{z} + \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} \log |\det \mathbf{J}_{zx}(\mathbf{z})| d\mathbf{z} \\ &= H_Z + \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [\log |\det \mathbf{J}_{zx}(\mathbf{z})|] \end{aligned} \quad (10)$$

Hence we have:

$$\begin{aligned} \text{KL}_\theta [\mu_Z \| q_Z] &= -H_Z + \log Z_X + \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [u(F_{zx}(\mathbf{z}; \theta))] - \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [\log |\det \mathbf{J}_{zx}(\mathbf{z}; \theta)|] \\ &= -H_X + \log Z_X + \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [u(F_{zx}(\mathbf{z}; \theta))] \\ &= -H_X + \log Z_X + \mathbb{E}_{\mathbf{x} \sim \mu_X(\mathbf{x}; \theta)} [u(\mathbf{x})] \\ &= \text{KL}_\theta [q_X \| \mu_X]. \end{aligned}$$

The loss function becomes a free energy:

$$\begin{aligned} J_{KL} &= \mathbb{E}_{\mathbf{z} \sim \mu_Z(\mathbf{z})} [u(F_{zx}(\mathbf{z}; \boldsymbol{\theta})) - \log |\det \mathbf{J}_{zx}(\mathbf{z}; \boldsymbol{\theta})|] \\ &= U - H_{ZX} \end{aligned}$$

with enthalpy  $U$  and transformation entropy:

$$H_{ZX} = H_X - H_Z$$

**Configuration KL divergence** Likewise, we can express the KL divergence in  $\mathbf{x}$  space where we compare the generated distributions with a Boltzmann weight. Using (3-4) and the Gaussian prior density (6), this KL-divergences evaluates as:

$$\begin{aligned} \text{KL}_{\boldsymbol{\theta}} [\mu_X \parallel q_X] &= H_X - \int \mu_X(\mathbf{x}) \log q_X(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \\ &= H_X - \int \mu_X(\mathbf{x}) [\log \mu_Z(F_{xz}(\mathbf{x}; \boldsymbol{\theta})) + \log |\mathbf{J}_{xz}(\mathbf{x}; \boldsymbol{\theta})|] d\mathbf{x}. \\ &= H_X + \log Z_Z + \mathbb{E}_{\mathbf{x} \sim \mu(\mathbf{x})} \left[ \frac{1}{\sigma^2} \|F_{xz}(\mathbf{x}; \boldsymbol{\theta})\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x}; \boldsymbol{\theta})| \right]. \end{aligned}$$

Although the constants  $H_X$  and  $Z_Z$  can be ignored during the training, this loss is difficult to evaluate because it needs to sample configurations according to  $\mu(\mathbf{x})$ , which is actually the problem we are trying to solve.

**Maximum Likelihood** However we can approximate the configuration KL divergence by starting from a sample  $\rho(\mathbf{x})$  and using the loss:

$$\begin{aligned} J_{ML} &= -\mathbb{E}_{\mathbf{x} \sim \rho(\mathbf{x})} [\log q_X(\mathbf{x}; \boldsymbol{\theta})] \\ &= \mathbb{E}_{\mathbf{x} \sim \rho(\mathbf{x})} \left[ \frac{1}{\sigma^2} \|F_{xz}(\mathbf{x}; \boldsymbol{\theta})\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x}; \boldsymbol{\theta})| \right] \end{aligned}$$

This loss is the negative log-likelihood, i.e. minimizing  $\text{LL}_{\boldsymbol{\theta}}$  corresponds to maximizing the likelihood of the sample  $\rho(\mathbf{x})$  in the Gaussian prior density.

**Symmetric divergence** The two KL divergences above can be naturally combined to the symmetric divergence

$$\text{KL}_{\text{sym}} = \frac{1}{2} \text{KL} [\mu_X \parallel q_X] + \frac{1}{2} \text{KL} [\mu_Z \parallel q_Z]$$

which corresponds, up to an additive constant, to the Jensen-Shannon divergence which uses the geometric mean of  $m = \sqrt{q_X q_Z}$  instead of the arithmetic mean.

**Reaction coordinate loss** In some applications we do not want to sample from the Boltzmann distribution but promote the sampling of high-energy states in a specific direction of configuration space, for example in order to compute a free energy profile along a predefined reaction coordinate  $R(\mathbf{x})$  (Fig. 2g). This is achieved by adding the reaction-coordinate (RC) loss to the minimization

problem:

$$\begin{aligned} J_{RC} &= \int p(R(\mathbf{x})) \log p(R(\mathbf{x})) dR(\mathbf{x}) \\ &= \mathbb{E}_{\mathbf{x} \sim q_X(\mathbf{x})} \log p(R(\mathbf{x})). \end{aligned}$$

To implement this loss, the function  $R$  is a user input, minimum and maximum bounds are given, and  $p(R(\mathbf{x}))$  is computed as a batchwise kernel density estimate along between the bounds.

### C. Systems and Hyperparameters

The “MD” simulations of the systems below are not using actual dynamics, but are emulated with Metropolis Monte Carlo with small local steps. In each step, a random vector from an isotropic Gaussian distribution with a system-dependent standard deviation  $\sigma_{\text{Metro}}$  is added to the present configuration. This proposed configuration is accepted or rejected with a standard Metropolis acceptance criterion.

All Boltzmann Generator networks are build of invertible blocks using NICE or RealNVP layers. Each block contains two such layers to make sure that all dimensions are subject to a nonlinear transformation, as described in SI Section A. Each configuration  $\mathbf{x}$  or latent vector  $\mathbf{z}$  is split into a channel of “even” and “odd” dimensions, defining the pairs  $(\mathbf{x}_1, \mathbf{x}_2)$  and  $(\mathbf{z}_1, \mathbf{z}_2)$ , respectively. To describe the network architecture used, we use  $N$ ,  $R$  and  $S$  to denote NICE block, RealNVP block and Scaling layer, respectively. A subscript is used to denote the number of repetitions of a motif, e.g.  $N_{10}$  are ten stacked NICE blocks, i.e. 20 layers total,  $(NR)_4$  are four repetitions of a NICE and a RealNVP block, i.e. 8 layers total.

**Double well** We define a two-dimensional toy model which is bistable in  $x$ -direction and harmonic in  $y$ -direction:

$$E(x, y) = \frac{1}{4}ax^4 - \frac{1}{2}bx^2 + cx + \frac{1}{2}dy^2 \quad (11)$$

with  $a = c = d = 1$  and  $b = 6$  – see Fig. 2 for the potential in  $x$ -direction. The system is simulated with a Metropolis step of  $\sigma_{\text{Metro}} = 0.1$ . To estimate the average time needed for a return trip between both states, we construct another systems with  $a = 0.25$  and  $b = 1.5$  that has the same position of minima and the same energy difference between them, but a much smaller barrier. For the “flat” systems frequent transitions between the two end-states are observed. The return-trip time of the original system is then estimated by  $t = t_{\text{flat}} \exp(B - B_{\text{flat}})$ , where  $B, B_{\text{flat}}$  are the energy barriers for the original and the “flat” system from either one of the two minima, and  $t, t_{\text{flat}}$  are the times taken for a round-trip between the states. This results in an estimate of  $t = 4 \cdot 10^6$  simulation steps for a return trip in the system shown in Fig. 2.

In Fig. 2 and Suppl. Figs. 6,7 we use NICE and RealNVP networks defined as below. Networks are trained in two steps: in a first ML phase we only minimize  $J_{\text{KL}}$ . Subsequently we minimize  $J_{\text{KL}} + J_{\text{ML}}$  with equal weights  $w_{\text{ML}} = w_{\text{KL}} = 1$ , unless otherwise noted.

Network	Epochs ML	Epochs KL
$N_{10}S$	200	1000
$R_{10}$	200	1000

**Bistable particle dimer in a Lennard-Jones fluid** Here we simulate two-dimensional system of a bistable particle dimer in a dense bath of  $n_s = 36$  solvent particles with Lennard-Jones repulsion. A similar system has been proposed in [14]. The configuration vector is defined by alternating  $x-$  and  $y-$  positions and starting with the two dimer particles:

$$\mathbf{x} = [\mathbf{x}_{1x}, \mathbf{x}_{1y}, \mathbf{x}_{2x}, \mathbf{x}_{2y}, \dots, \mathbf{x}_{(n_s+2)x}, \mathbf{x}_{(n_s+2)y}] .$$

Defining the dimer distance  $d = \|\mathbf{x}_1 - \mathbf{x}_2\|$ , and the Heavyside step function  $h$ , we use the potential energy:

$$\begin{aligned} U(\mathbf{x}) = & k_d (\mathbf{x}_{1x} + \mathbf{x}_{2x})^2 + k_d \mathbf{x}_{1y}^2 + k_d \mathbf{x}_{2y}^2 \\ & + \frac{1}{4} a(d - d_0)^4 - \frac{1}{2} b(d - d_0)^2 + c(d - d_0)^4 \\ & + \sum_{i=1}^{n+2} h(-\mathbf{x}_{ix} - l_{\text{box}}) k_{\text{box}} (-\mathbf{x}_{ix} - l_{\text{box}})^2 + \sum_{i=1}^{n+2} h(\mathbf{x}_{ix} - l_{\text{box}}) k_{\text{box}} (\mathbf{x}_{ix} - l_{\text{box}})^2 \\ & + \sum_{i=1}^{n+2} h(-\mathbf{x}_{iy} - l_{\text{box}}) k_{\text{box}} (-\mathbf{x}_{iy} - l_{\text{box}})^2 + \sum_{i=1}^{n+2} h(\mathbf{x}_{iy} - l_{\text{box}}) k_{\text{box}} (\mathbf{x}_{iy} - l_{\text{box}})^2 \\ & + \epsilon \sum_{i=1}^{n+1} \sum_{j=i+1, j \neq 2}^{n+2} \left( \frac{\sigma}{\|\mathbf{x}_i - \mathbf{x}_j\|} \right)^{12} \end{aligned}$$

where the five rows correspond to: (1) Constraints for the center and the  $y$ -position of the particle dimer, (2) particle dimer interaction, (3,4) box constraints in  $x-$  and  $y$ -direction, (5) particle repulsion. The following parameter values were used (all in reduced units):

Parameter	$\epsilon$	$\sigma$	$k_d$	$d_0$	$a$	$b$	$c$	$l_{\text{box}}$	$k_{\text{box}}$
Value	1.0	1.1	20.0	1.5	25.0	10.0	-0.5	3.0	100.0

To initialize training, we run Metropolis Monte Carlo simulations with a Metropolis step length of  $\sigma_{\text{Metro}} = 0.02\sqrt{\tau}$ , where  $\tau$  is the relative temperature. To estimate the time taken for a return-trip between open and closed dimer states, we take the same approach as for the double-well system above: We conduct a simulation with  $10^6$  simulation steps for a system with maximally flattened energy ( $a = 2.5$  and  $b = 1.0$ ). Still no transition from closed to open states occur, we thus estimate the *lower bound* for the return trip to be  $t = 10^6 \exp(B - B_{\text{flat}}) \approx 1.2 \cdot 10^{12}$  where  $B, B_{\text{flat}}$  are the intrinsic barrier heights for the unchanged and flattened system.

For validation of the free energy profiles predicted in Fig. 3i, we perform Umbrella Sampling simulations [17] for each relative temperature (0.5, 1.0, 2.0) using 35 Umbrella potentials on the dimer distance between values of 0.5 and 2.5 and with a force constant of 500 (reduced units).

For initializing the training by example (ML),  $10^5$  simulation steps are stored for the “open” and “closed” dimer states. No transitions between these states occur in the simulations. In order to not have to learn the permutational invariance of the diffusing solvent particles from the data, we remove this invariance by relabeling solvent particles using the Hungarian algorithm [9].

Boltzmann Generator training was done with  $w_{KL} = 1$  and decreasing  $w_{ML}$  according to the following schedule:

Epochs	20	200	300	300	1000	2000
$w_{ML}$	1	100	100	100	20	<i>0.01</i>
$w_{KL}$	0	1	1	1	1	1
$w_{RC}$	0	1	5	10	10	<i>10</i>

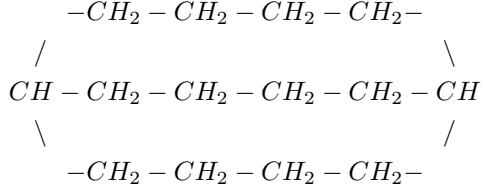
The italic values in the last row were treated as variable hyperparameters. We then did a hyperparameter search as indicated in the table below. The hyperparameters were chosen by minimizing the estimator variance for the free energy profile along dimer distance  $d$ . Each trained network makes predictions for the free energy profile shown in Fig. 3i. Using bootstrapping the standard error over all free energies along the profile between  $d = [0.5, 2.5]$  are computed, resulting in  $(\epsilon_{0.5}, \epsilon_{1.0}, \epsilon_{2.0})$  for the three temperatures and  $\bar{\epsilon} = \sqrt{\epsilon_{0.5}^2 + \epsilon_{1.0}^2 + \epsilon_{2.0}^2}$  as a total estimator error.

Architecture	$nl_{layers}$	$nl_{hidden}$	$w_{ML}$	$w_{RC}$	$\epsilon_{0.5}$	$\epsilon_1$	$\epsilon_2$	$\sqrt{\sum \epsilon^2}$
R <sub>8</sub>	4	200	0.1	10.0	1.62	2.07	2.04	3.33
R <sub>4</sub>	.	.	.	.	2.23	1.83	1.53	3.27
R <sub>6</sub>	.	.	.	.	1.69	1.64	2.29	3.28
R <sub>12</sub>	.	.	.	.	1.49	1.85	2.0	3.10
(RN) <sub>2</sub>	.	.	.	.	2.01	1.59	2.56	3.62
(RN) <sub>7</sub>	.	.	.	.	1.76	2.29	1.39	3.20
(RN) <sub>12</sub>	.	.	.	.	1.44	1.66	1.99	2.96
R <sub>8</sub>	3	200	0.1	10.0	1.51	1.97	1.64	2.97
R <sub>4</sub>	.	.	.	.	1.41	1.59	1.78	2.77
R <sub>6</sub>	.	.	.	.	1.49	1.73	1.76	2.88
R <sub>12</sub>	.	.	.	.	1.84	1.28	2.24	3.17
(RN) <sub>2</sub>	.	.	.	.	1.84	1.29	2.25	3.18
(RN) <sub>7</sub>	.	.	.	.	1.65	1.72	2.07	3.16
(RN) <sub>12</sub>	.	.	.	.	2.87	1.80	1.58	3.74
R <sub>8</sub>	2	.	.	.	1.85	1.58	2.50	3.48
.	4	.	.	.	1.69	1.51	1.52	2.73
.	3	50	.	.	1.32	1.71	2.11	3.02
.	.	100	.	.	2.85	2.05	2.16	4.12
.	.	<b>200</b>	<b>0.01</b>	.	<b>1.58</b>	<b>1.33</b>	<b>1.33</b>	<b>2.45</b>
.	.	.	1.0	.	1.87	1.93	1.63	3.15
.	.	.	0.1	1.0	1.66	1.83	1.75	3.02
.	.	.	.	5.0	1.73	1.72	1.81	3.03
.	.	.	.	20.0	1.88	2.06	1.84	3.34

Table 2: Hyperparameter selection for the particle dimer. In the architecture,  $R$  corresponds to a RealNVP and  $N$  to a NICE double layer with channel swaps (Fig. 1). The subscript indicates the number of repetitions, e.g. R<sub>4</sub> = RRRR, corresponding to eight single layers. All nonlinear transformations ( $T$ ,  $S$ ) used dense networks with tanh activation using the given number of layers ( $nl_{layers}$ ) and hidden nodes ( $nl_{hidden}$ ). All networks were trained on the following range of relative temperatures:  $\tau \in [0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4]$  and used  $w_{KL} = 1.0$ .

**Hydrocarbons** A simple molecular mechanics model including bond, angle, torsion and Lennard-Jones interactions (between all pairs) was implemented in TensorFlow. We use the parameters in [13] to modeling the alkanes shown in Fig. 4: (1) cyclononane  $C_9H_{18}$ , i.e. a ring of nine  $CH_2$  groups, and

(2) bicyclo[4.4.4]tetradecane  $C_{14}H_{26}$ , connected as follows:



To generate some initial structures for training by example, we conducted a short (14,000 steps) replica-exchange simulation with relative temperatures  $\tau = (1, 1.5, 2, 2.5, 3, 3.5, 4)$  where 300 Kelvin is the standard temperature. We used Metropolis-Monte Carlo step length  $\sigma_{\text{Metro}} = 0.01 \text{ \AA}$ . The main aim of this simulation is to equilibrate the structure which is started from a random placement of atoms in space. The first 10,000 steps were discarded and the last 4,000 steps were retained for training.

The Boltzmann Generator is trained using iterative discovery. In each of 50 iterations, we use 4,000 configurations for training by example, and we train using the following schedule:

	First Iteration			Next Iterations
Epochs	300	300	300	300
$w_{ML}$	1	1	1	1
$w_{KL}$	0	0.01	0.1	0.1

At the end of each iteration, we use the Boltzmann Generator to sample  $10^6$  structures and make a list of all rotamer conformations that have been generated. A rotamer is defined by discretizing each torsion as follows:

Rotamer	1	2	3
Angle	$-120^\circ \dots 0^\circ$	$0^\circ \dots 120^\circ$	$120^\circ \dots -120^\circ$

And the rotamer conformation is the combination of all rotamers, e.g. “123211211”. We remove permutational symmetries, in cycloalkanes this is achieved by setting all rotamer conformations equal that can be transformed by mirroring or cyclical permutation. After listing all conformations, we resample a list of 4,000 configurations, but now each rotamer conformation found has an equal probability of being sampled. This list is inserted into the next iteration for training by example.

## D. Additional Figures

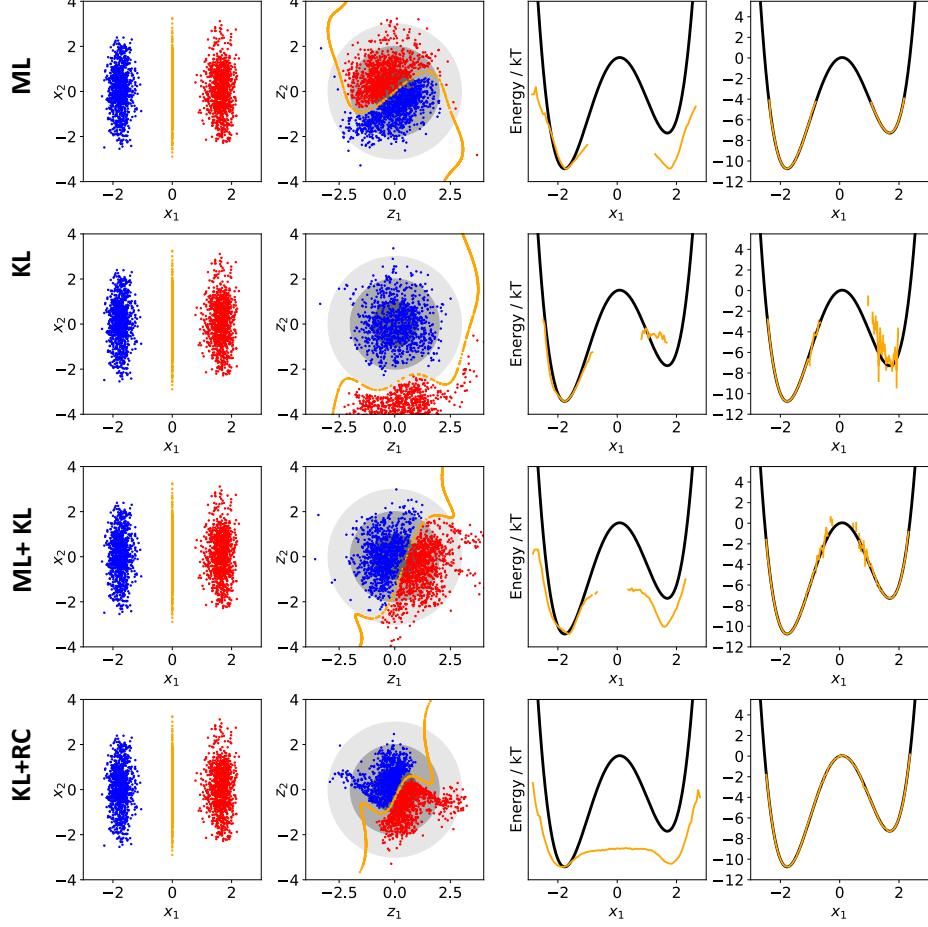


Figure 6: Different methods for training Boltzmann Generators with NICE layers using the double well example shown in Fig. 2. Columns show: (1) distribution in configuration space  $\mathbf{x}$ , (2) distribution in latent space  $\mathbf{z}$ , (3) free energy of Boltzmann Generator output  $p_X(\mathbf{x})$  along  $x_1$ , (4) free energy after reweighting. Training uses 200 epochs of ML and then 500 epochs of the method given in the rows, using equal weights for these modes. Training by example (ML) only reproduces the distribution of the training data, which can be reweighted to the Boltzmann distribution in this low-dimensional example but reweighting fails for high-dimensional examples. Training by energy (KL) only tends to collapse to one metastable state. ML+KL combined samples closer to the Boltzmann distribution than ML and avoids metastable state collapse, but samples high-energy transition states with low probability. KL+RC performs best in this example.

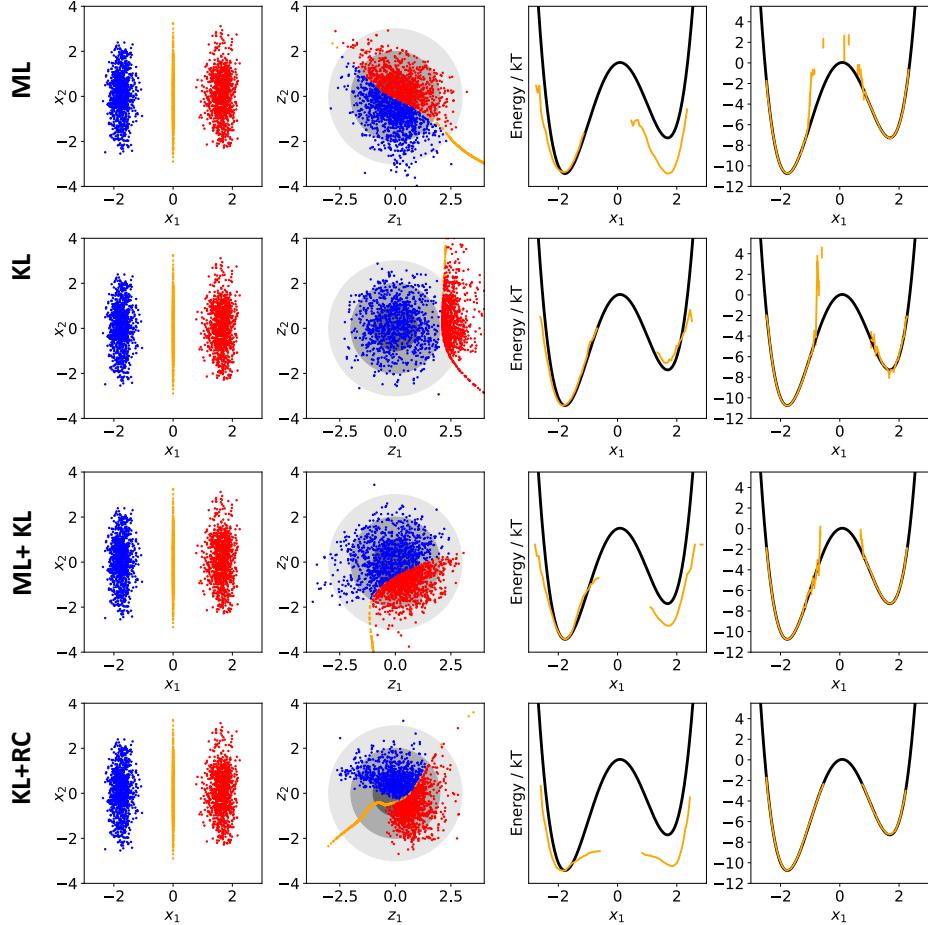


Figure 7: Same as Fig. 6 but for Boltzmann Generators using RealNVP layers.