Supplementary Material for "Transforming Bell's Inequalities into State Classifiers with Machine Learning"

Contents

I. POSITIVE PARTIAL TRANSPOSE (PPT)

In this section, we will introduce Positive Partial Transpose (PPT) criterion [1]. It is an analytic algorithm to determine whether an ensemble ρ is entangled or not. As an example, PPT criterion will then be applied to a specific ensemble, which we have used to demonstrate our machine learning method in the main text.

We can expand any ensemble ρ of 2 quantum systems A and B in a given product basis as

$$\rho = \sum_{i,j,k,l} \rho_{kl}^{ij} \left|i\right\rangle \left\langle j\right| \otimes \left|k\right\rangle \left\langle l\right| \ . \tag{1}$$

The partial transposition (with respect to the second party) is

$$\rho^{T_B} = \sum_{i,j,m,n} \rho_{mn}^{ij} \left|i\right\rangle \left\langle j\right| \otimes \left(\left|m\right\rangle \left\langle n\right|\right)^T = \sum_{i,j,m,n} \rho_{mn}^{ij} \left|i\right\rangle \left\langle j\right| \otimes \left|n\right\rangle \left\langle m\right| \ . \tag{2}$$

For 2-qubit system, $i, j, m, n \in 0, 1$. The state in n-qubit system can also be regarded as bipartite state. For example, system A consists of 1 qubit and system B includes $n-1$ qubits.

According to PPT criterion, if ρ is a bipartite separable state, the minimum eigenvalue of partial transposition of ρ is non-negative. The inverse proposition is also true for 2-qubit system.

Then define the entangled states we care about.

$$\left|\psi_{\theta,\phi}\right\rangle = \cos\frac{\theta}{2}\left|00\right\rangle + e^{i\phi}\sin\frac{\theta}{2}\left|11\right\rangle \tag{3}$$

$$\rho_{\theta,\phi} = p\left|\psi_{\theta,\phi}\right\rangle\left\langle\psi_{\theta,\phi}\right| + (1-p)\, I/4 \ , \tag{4}$$

where $0 \le \theta \le \pi$, $0 \le \phi < 2\pi$ and $0 \le p \le 1$. I is identity matrix. For convenience to discuss later, we define

$$\left|0\right\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \left|1\right\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \ , \tag{5}$$

and we have

$$\rho_{\theta,\phi} = \begin{bmatrix} p\cos^2\frac{\theta}{2} + \frac{1-p}{4} & 0 & 0 & pe^{-i\phi}\cos\frac{\theta}{2}\sin\frac{\theta}{2} \\ 0 & \frac{1-p}{4} & 0 & 0 \\ 0 & 0 & \frac{1-p}{4} & 0 \\ pe^{i\phi}\cos\frac{\theta}{2}\sin\frac{\theta}{2} & 0 & 0 & p\sin^2\frac{\theta}{2} + \frac{1-p}{4} \end{bmatrix} \ , \tag{6}$$

$$\rho_{\theta,\phi}^{T_B} = \begin{bmatrix} p\cos^2\frac{\theta}{2} + \frac{1-p}{4} & 0 & 0 & 0 \\ 0 & \frac{1-p}{4} & pe^{-i\phi}\cos\frac{\theta}{2}\sin\frac{\theta}{2} & 0 \\ 0 & pe^{i\phi}\cos\frac{\theta}{2}\sin\frac{\theta}{2} & \frac{1-p}{4} & 0 \\ 0 & 0 & 0 & p\sin^2\frac{\theta}{2} + \frac{1-p}{4} \end{bmatrix} \ . \tag{7}$$

The 4 eigenvalues of $\rho_{\theta,\phi}^{T_B}$ are $\lambda_1 = p\cos^2\frac{\theta}{2} + \frac{1-p}{4}$, $\lambda_2 = p\sin^2\frac{\theta}{2} + \frac{1-p}{4}$, $\lambda_3 = \frac{1-p}{4} + p\cos\frac{\theta}{2}\sin\frac{\theta}{2}$ and

$$\lambda_{\min} = \frac{1-p}{4} - p\cos\frac{\theta}{2}\sin\frac{\theta}{2} \ . \tag{8}$$

$\lambda_{\min}$ is the minimum eigenvalue and thus $\lambda_{\min} < 0$ iff $\rho_{\theta,\phi}$ is entangled. If $\theta = \pi/2$, the condition becomes $p > 1/3$.

In general, PPT depends $4^n - 1$ measurement outcomes in n-qubit system (i.e. tomography), which is resource-consuming. The computational resources depend on the how to evaluating the minimum eigenvalue of a $D \times D$ matrix, here $D = 2^n$.

II. CHSH INEQUALITY

CHSH (stands for John Clauser, Michael Horne, Abner Shimony, and Richard Holt) inequality, which is one of the most well-known Bell inequalities, implies the existence of local hidden variables if it is always true. In other words,

the violation of CHSH inequality is a sufficient condition to ensure the state to be entangled. In this section, we apply CHSH inequality to the states of the first experiment in the main text, and show its limitation for entanglement detection.

Define CHSH operator Π_{CHSH} by assembling the elements of standard CHSH inequality.

$$\Pi_{\text{CHSH}} = \sigma_z \frac{\sigma_x - \sigma_z}{\sqrt{2}} - \sigma_z \frac{\sigma_x + \sigma_z}{\sqrt{2}} + \sigma_x \frac{\sigma_x - \sigma_z}{\sqrt{2}} + \sigma_x \frac{\sigma_x + \sigma_z}{\sqrt{2}} \,. \tag{9}$$

Thus $-2 \leq \text{Tr}(\rho\Pi_{\text{CHSH}}) \leq 2$, which for all separable ensemble ρ predicted by CHSH inequality [2]. However, for all the entangled states, the upper bound of $|\text{Tr}(\rho\Pi_{\text{CHSH}})|$ is $2\sqrt{2}$ which was derived from Tsirelson [3].

It is easy to validate the results below.

$$\langle 00| \Pi_{\text{CHSH}} |00\rangle = \langle 00| - \sqrt{2}\sigma_z\sigma_z |00\rangle = -\sqrt{2} \tag{10}$$
$$\langle 11| \Pi_{\text{CHSH}} |11\rangle = \langle 11| - \sqrt{2}\sigma_z\sigma_z |11\rangle = -\sqrt{2} \tag{11}$$
$$\langle 00| \Pi_{\text{CHSH}} |11\rangle = \langle 00| \sqrt{2}\sigma_x\sigma_x |11\rangle = \sqrt{2} \tag{12}$$
$$\langle 00| \Pi_{\text{CHSH}} |11\rangle = \langle 00| \sqrt{2}\sigma_x\sigma_x |11\rangle = \sqrt{2} \,. \tag{13}$$

Then we obtain

$$\begin{aligned} & \langle \psi_{\theta,\phi}| \Pi_{\text{CHSH}} |\psi_{\theta,\phi}\rangle \\ = \; & (\cos\frac{\theta}{2} \langle 00| + e^{-i\phi} \sin\frac{\theta}{2} \langle 11|) \Pi_{\text{CHSH}} (\cos\frac{\theta}{2} |00\rangle + e^{i\phi} \sin\frac{\theta}{2} |11\rangle) \\ = \; & -\sqrt{2} + 2\sqrt{2} \cos\frac{\theta}{2} \sin\frac{\theta}{2} \cos\phi \\ = \; & \sqrt{2}(\sin\theta\cos\phi - 1) \\ \geq \; & -2\sqrt{2} \,. \end{aligned} \tag{14}$$

If $\theta = \pi/2$, $\phi = \pi$, $|\psi_{\theta,\phi}\rangle$ becomes $\frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)$ which maximumly violates CHSH inequality. Since $\text{Tr}(\Pi_{\text{CHSH}}) = 0$, we get

$$\text{Tr}\left(\rho_{\theta,\phi}\Pi_{\text{CHSH}}\right) = \text{Tr}\left(p \langle \psi_{\theta,\phi}| \Pi_{\text{CHSH}} |\psi_{\theta,\phi}\rangle\right) = \sqrt{2}p(\sin\theta\cos\phi - 1) = -2\sqrt{2}p \,. \tag{15}$$

Thus for $\frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)$, $\rho_{\theta,\phi}$ will not violate CHSH inequality iff $p \leq 1/\sqrt{2}$. However, according to PPT we discussed in Section I, all the ensembles with $p > 1/3$ are entangled. Therefore some entangled ensembles can not be detected by this CHSH inequality.

III. ENTANGLEMENT WITNESS

Similar to CHSH inequality, entanglement witness is an another analytic tool to detect entanglement. (Bell inequality is just a specific case of witness.) More specifically speaking, an observable $\mathcal{W}$ can be regarded as an entanglement witness iff

$$\text{Tr}(\rho\mathcal{W}) < 0 \quad \text{for at least 1 entangled ensemble } \rho \,, \tag{16}$$
$$\text{Tr}(\rho\mathcal{W}) \geq 0 \quad \text{if } \rho \text{ is separable} \,. \tag{17}$$

We say $\mathcal{W}$ detect ρ if $\text{Tr}(\rho\mathcal{W}) < 0$. As an example, to detect an entangled state $|\psi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, we construct

$$\mathcal{W}_+ = \frac{I}{2} - |\psi_+\rangle \langle\psi_+| = \frac{1}{2} \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} \,. \tag{18}$$

It is easy to verify that $\langle\psi_+| \mathcal{W}_+ |\psi_+\rangle < 0$, thus $|\psi_+\rangle$ is detected by $\mathcal{W}_+$. it has been shown [4] that the witness must be decomposed at least in 3 measurements per party, while CHSH inequality only needs 2. However, for more general ensemble $\rho_{\theta,\phi}$ with unknown p, θ and ϕ, $\mathcal{W}_+$ can also detect some of entangled ensembles. Because

$$\text{Tr}\left(\rho_{\theta,\phi}\mathcal{W}_+\right) = \frac{1-p}{4} - p\cos\frac{\theta}{2}\sin\frac{\theta}{2}\cos\phi \,. \tag{19}$$

Compared with 8, if $\phi = 0$, all the entangled ensembles can be detected by $\mathcal{W}_+$ perfectly whatever p and θ is. However, if we do not know anything about ϕ, this method is not satisfactory to detect all the entanglement states.

IV. MULTI-QUBIT SYSTEM

As we have discussed in the main text, in order to detect unknown entangled states, tomographic information is necessary no matter what algorithm is implemented. However, the size of measurement outcomes grows exponentially with qubit number increasing. Therefore, general entanglement detector can not be scalable for multi-qubit systems.

In this section, we focus on a typical n-qubit entanglement state, called Greenberger-Horne-Zeilinger (GHZ) state, i.e.

$$|\psi_{GHZ}\rangle = (|0\rangle^{\otimes n} + |1\rangle^{\otimes n})/\sqrt{2} \tag{20}$$

Note that the reduced density matrix of any $n-1$ particles, i.e. $\frac{1}{2}|0\rangle^{\otimes n-1}\langle 0|^{\otimes n-1} + \frac{1}{2}|1\rangle^{\otimes n-1}\langle 1|^{\otimes n-1}$, is fully separable. Therefore in order to detect such an entangled state by PPT criterion, fully information of n-qubit system must be obtained.

Therefore, it is reasonable to train our ANN architecture using only partial information to distinguish GHZ-type states from fully separable ones. The data are generated individually, therefore "label problem" does not exist. For more general cases, some numerical methods like semidefinite programming could be helpful to label data. Here fully separable states are generated according to its definition

$$\rho_{\text{sep}} = \sum_{i=1}^{D} p_i \, \rho_i^1 \otimes \rho_i^2 \otimes \cdots \otimes \rho_i^n \tag{21}$$

We set $D = 7$ in this task. $p_i \in [0,1]$ are sampled uniformly and normalized to ensure $\sum_{i=1}^{D} p_i = 1$. ρ_i is sampled directly by the code RandomDensityMatrix [5].

GHZ-type states are generated through $U_1 \otimes U_2 \cdots U_n |\psi_{GHZ}\rangle$ (the so-called LOCC) or $\sigma_1 \otimes \sigma_2 \cdots \sigma_n |\psi_{GHZ}\rangle$ (and normalizing, i.e. SLOCC). Here $U_{1,2,\cdots n}$ and $\sigma_{1,2,\cdots n}$ are random unitary and random invertible matrices on each qubit, which have been entirely discussed in the main text. More various types of entangled states are generated by SLOCC, therefore in order to identifying them, more features are potentially necessary.

The size of input layer (feature) in our predictors is n for LOCC or $2n$ for SLOCC [6], linearly increases with qubit number. Each element of features is constitutive of n random Pauli matrices, i.e. $\langle \boldsymbol{O^1 O^2 \cdots O^n} \rangle$. Therefore n^2 (for LOCC) or $2n^2$ (for SLOCC) local observables are necessary to be used for our models.

The results are shown in Fig. 1. We find that our model performs satisfactory if given enough hidden units and features. For many-body system, the test errors have little correlation with qubit number.

The procedures are confined to the problems in training set only. For 8-qubit system, the time of generating raw data are far longer than that of training a predictor. Fortunately, in future, we can imagine that these resource-demanding tasks can be achieved by a few super (quantum or classical) computers. Once a predictor is constructed, any small laboratory can makes use of it by measuring only a relatively small amount of features of a testing quantum states. In this sense, the task of solving the computational problems can be shared by users with different computational powers.

V. BISEPARABLE THREE-QUBIT SYSTEM

In this section, we implement ANN to identify different biseparable states in 3-qubit system. As we have mentioned in the main text, a three-qubit quantum state is called biseparable, if two of the qubits are entangled with each other but not with the third one. The corresponding density matrices are denoted as

$$\{\, \rho_{A|BC}, \quad \rho_{B|AC}, \quad \rho_{C|AB} \,\}, \tag{22}$$

and their convex combination, i.e.,

$$\lambda_1 \, \rho_{A|BC} + \lambda_2 \, \rho_{B|AC} + \lambda_3 \, \rho_{C|AB} \tag{23}$$

for $0 \leq \lambda_1, \lambda_2, \lambda_3 \leq 1$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Of course, these sets of states include fully-separable states as a special case. A system is called fully entangled (or genuine tripartite entangled for pure states) [7] if it is neither biseparable nor fully-separable.

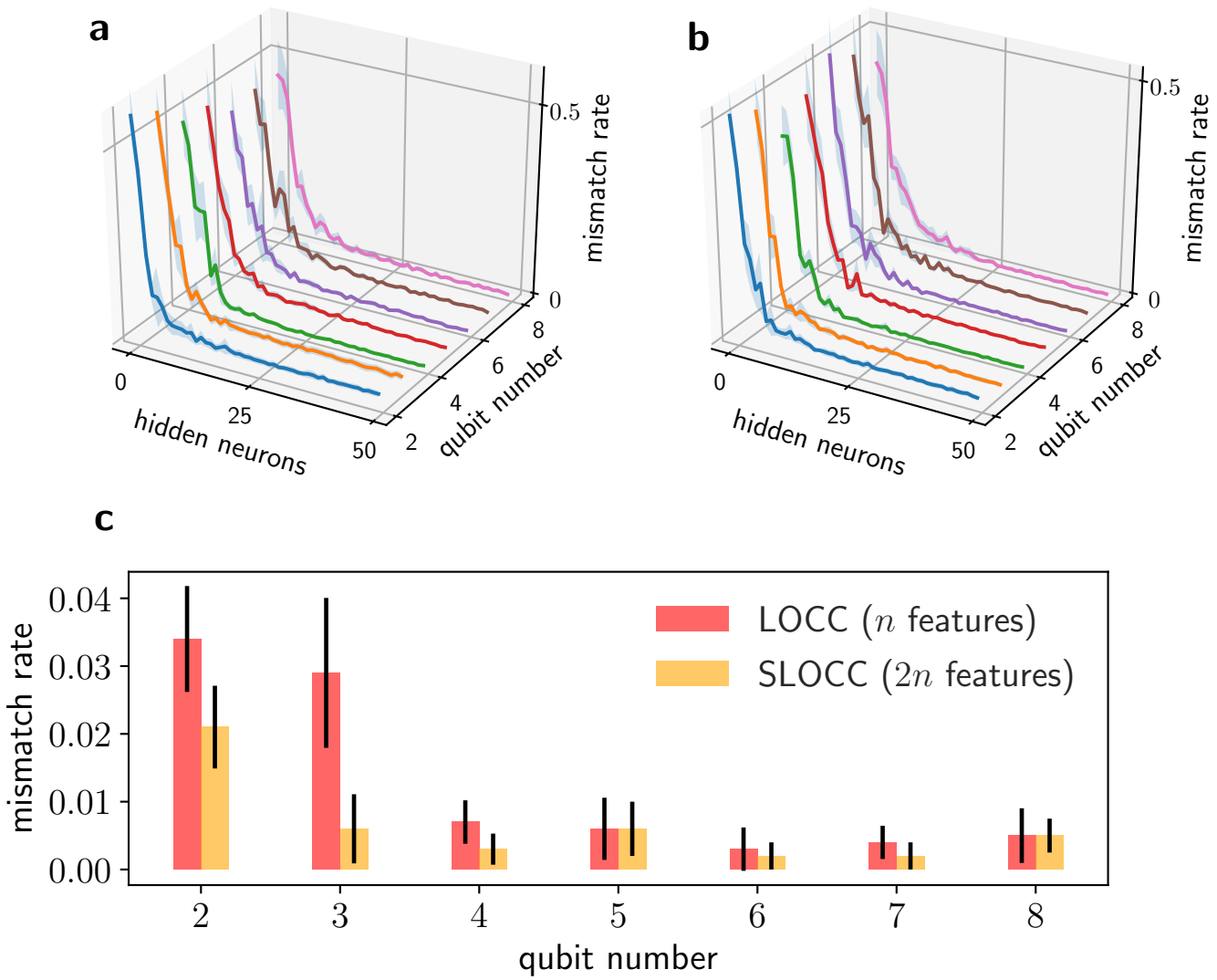


FIG. 1: **Test error (mismatch rate) for identifying GHZ-type states and fully separable states.** The data are generated according to **(a)** LOCC or **(b)** SLOCC . 10 different sets of both features and states are generated randomly for every qubit number. each set includes only 500 GHZ-type states and 500 fully separable states. 90% of states are used for training, others for testing. $\mu \pm \sigma/2$ are depicted, here μ is the mean value of 10 errors and σ is their standard deviation. **(c)** Test error with 49 hidden neurons.

In the following, we shall focus on the quantum ensemble of the following form (see Fig. 2**(a,b)**),

$$\rho = p\,\rho_{\rm bs} + (1-p)\,\rho_{\rm sep}\,, \tag{24}$$

where $\rho_{\rm bs} = |\psi_{\rm bs}\rangle\langle\psi_{\rm bs}| \otimes \rho_{\rm rand}$ belongs to one of the three biseparable states in Eq. (22), the value of $p \in [0,1]$ is generated uniformly, and $\rho_{\rm sep}$ is a random fully-separable states defined in Eq. 21 with $D = 17$ (same with the 3-qubit system in the main text). We apply $\rho_{\rm sep}$ rather than $I/8$ here because we want to cover all fully-separable states in our training data. The set of pure states $|\psi_{\rm bs}\rangle$ is generated by random vectors as follows. A random vector is defined by a vector,

$$\mathbf{v}_{\rm rand}(d) = \{v_1, v_2, v_3, ...\}\,, \tag{25}$$

containing d elements v_i from the Gaussian distribution with a zero mean and unit variance. Here the state $|\psi_{\rm bs}\rangle$ is obtained by $\mathbf{v}_{\rm rand}(4)$, followed by multiplying a normalization factor.

In this case, we can still label the entanglement of the quantum states by the use of the PPT criterion. For example, if the state ρ is in $\rho_{A|BC}$, then we can trace out the system A in the total density matrix of the form given in Eq. (24),

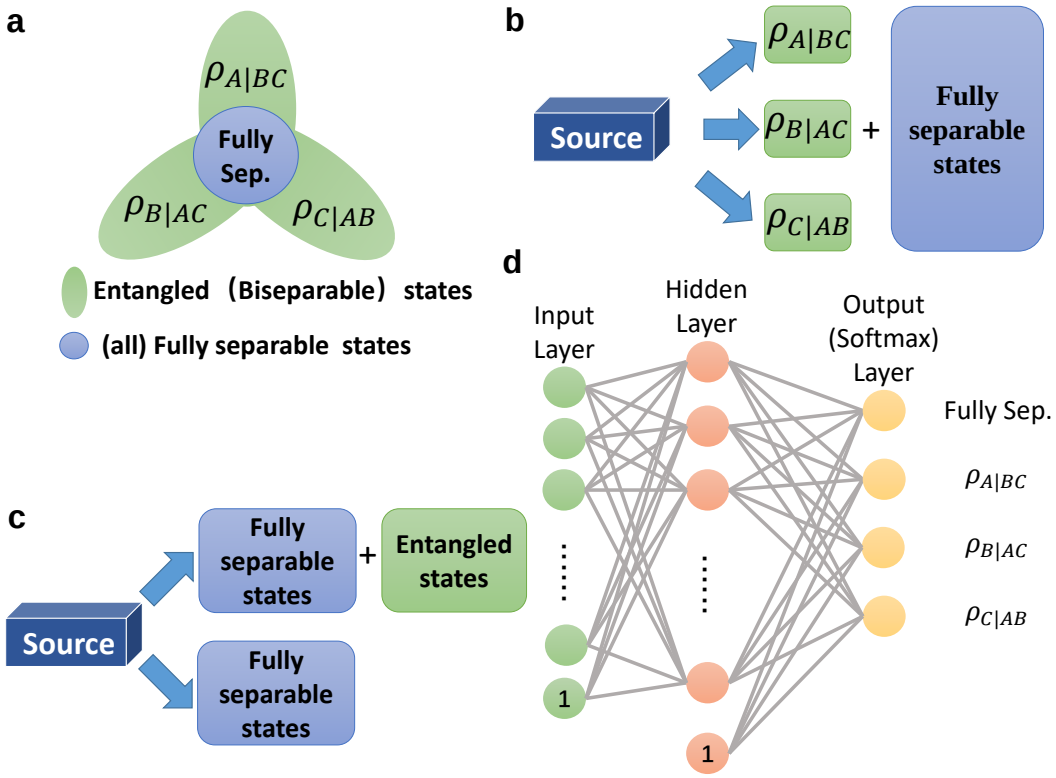


FIG. 2: **Multi(3 and 4)-qubit system. (a)** The relationship between 3 types of biseparable quantum states and fully-separable states. Fully separable state is just a specific case of biseparable state. **(b)** The generation process of 3-qubit biseparable states. We generated 200,000 states for each biseparable type and combined them with fully separable states. 90% of them are used for training and others for testing. By PPT criterion, we find that the population of fully-separable states is about 45% for 600,000 data points. **(c)** The generation process of 4-qubit system. One type of random states is assumed to be fully-separable, while the other is random pure entangled state and mixed with random fully separable states. 170,000 states are generated for each type. 90% of them are used for training and others for testing. **(d)** The ANN to distinguish multi-class quantum states. We apply the model to detect the entanglement of 3-qubit system. The output result is encoded on the Softmax layer $[c_0, c_1, c_2, c_3]^T (\sum_{i=0}^{3} c_i = 1)$ rather than single neuron because we want distinguish 4 types of states rather than 2, each element can be interpreted as the probability of each possible state. For example, if the output is $[0.7, 0.1, 0.1, 0.1]^T$, it can be interpreted as "the state is fully separable by 70% probability."

then apply the PPT criterion to the reduced density matrix of BC. In this way, we challenge our Bell-like predictors to classify four types of states, including 3 types of biseparable entangled states and the fully-separable states.

The results are shown in Fig. 3. Here we try to distinguish 4 types of different states. The task is expected to be more difficult and more resources-consuming than 2 types therefore the error is reasonable to be larger. In our machine learning method, we adopted the elements of Mermin inequality as input (4 features) to train our Bell-like predictor $\text{Bell}_{\text{ml}}(3, 4, x)$, and similarly, for the Svetlichny inequality, we constructed $\text{Bell}_{\text{ml}}(3, 8, x)$. In the cases discussed previously, the sigmoid function was employed for binary classification of the quantum states. Here the output layer is obtained by the Softmax function [8], which can be applied to multi-state classification (Fig. 2**d**).

The mismatch rate of the machine learning method is shown in Fig. 3. It is indicated that if we just use the same features as in Mermin $\text{Bell}_{\text{ml}}(3, 4, x)$ and Svetlichny $\text{Bell}_{\text{ml}}(3, 8, x)$ inequalities, the performance is not satisfactory. The mismatch rate cannot get much improved by increasing the number of neurons in the hidden layer. However, the performance of machine learning method can get significantly improved by putting 3 groups of features from CHSH inequalities for every pair of qubits, which gives a new $\text{Bell}_{\text{ml}}(3, 12, x)$ predictor. More improvement can be obtained by adding features, as $\text{Bell}_{\text{ml}}(3, 26, x)$ and tomographic predictors perform in this task. More details about the features are depicted in TABLE 1.

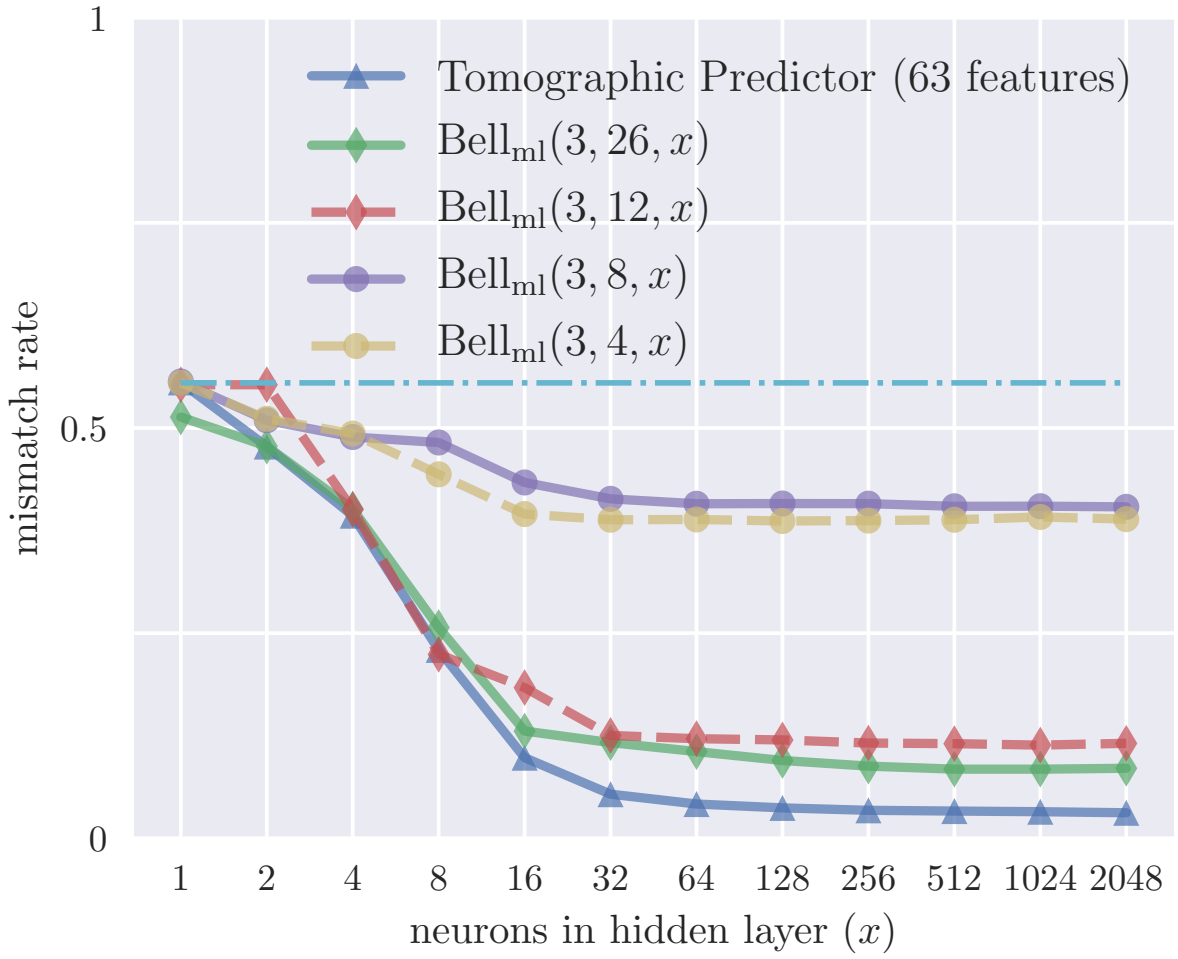


FIG. 3: **Biseparable states distinguished via Bell-like predictors.** 5 types of predictors applied on the biseparable ensembles generated according to Fig. 2**(a,b)**. The mismatch rate of our predictor decreases with the number of hidden layer neurons growing. The predictors that transformed from the Bell's inequalities that only detect fully-entangled states, such as $\mathrm{Bell}_{\mathrm{ml}}(3, 4, x)$ and $\mathrm{Bell}_{\mathrm{ml}}(3, 8, x)$, are not satisfying to distinguish biseparable (entangled) states from fully-separable states in our task. It can be significant improved by, the predictor with 3 groups of CHSH features, i.e $\mathrm{Bell}_{\mathrm{ml}}(3, 12, x)$. The horizontal dashed line represents the minimum mismatch rate for random guess, i.e. identify all the instances as fully-separable states.

VI. FOUR-QUBIT SYSTEM WITH INCORRECT LABEL

Finally, we apply our machine learning method to the case of a four-qubit system to classify two special classes of quantum states (Fig. 2**c**). The blue class of ensembles is fully-separable ρ_{sep} and the green is assumed to be the form

$$\rho_{\mathrm{mix}} = p\,|\psi_{\mathrm{rand}}\rangle\,\langle\psi_{\mathrm{rand}}| + (1-p)\rho_{\mathrm{sep}}\,, \tag{26}$$

where $|\psi_{\mathrm{rand}}\rangle$ is a totally random pure state (defined in Eq. (25) with $d = 16$) and $p \in [p_{\mathrm{min}}, 1]$ is a uniformly random variable. Note that we need to set a minimum value for p_{min}; it is because if $p = 0$, then $\rho_{\mathrm{mix}} = \rho_{\mathrm{sep}}$, making the two sets of states identical. ρ_{sep} is defined in Eq. 21 with $D = 7$

Recall that the PPT criterion identifies entangled state whenever the eigenvalue of the corresponding matrix, after partial transpose, becomes negative. Here we find numerically that when $p_{\mathrm{min}} > 0.1$, all of the instances in ρ_{mix} are entangled states. The scenarios is depicted in Fig. 4**(a-c)**.

As shown in Fig. 4**d**, we studied the performance of a class of Bell-like predictors, namely $\mathrm{Bell}_{\mathrm{ml}}(4, 80, x)$, where the neuron number in hidden layer is taken from 1 to 15, i.e., $x = 1, 2, \cdots 15$.

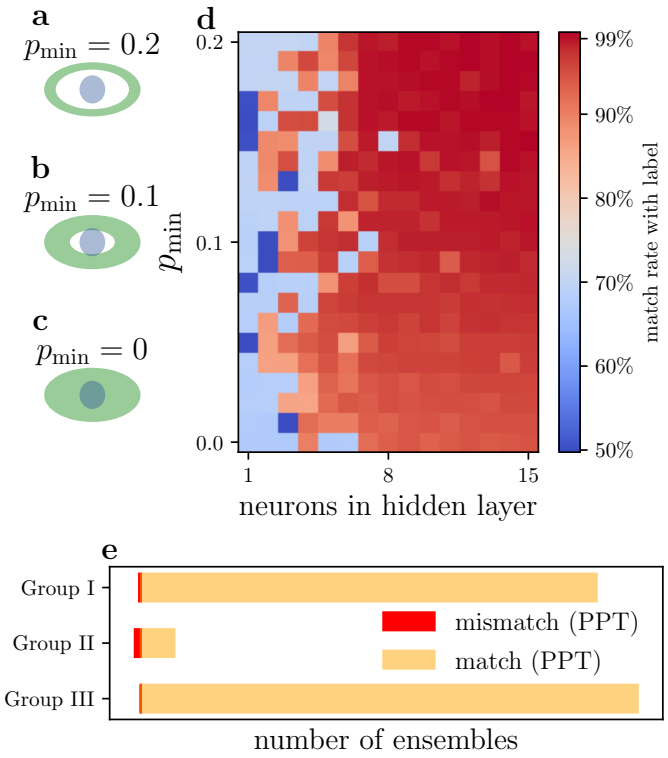


FIG. 4: **Quantum states of 4-qubit system distinguished by Bell-like(d) and tomographic(e) predictors. (a-c)** The diagram of states generated by two different ways (Fig. 2**c**). Blue represents fully-separable states and green are entangled states with p larger than $p_{\min}$ (Eq. 26). **(d)** The match rate of label and prediction by $\text{Bell}_{\text{ml}}(4, 80, x)$ with $p_{\min}$ (y-axis) and number of hidden layer neurons (x-axis) increasing. Here a fraction of states (**Group II**) are deliberately labelled inconsistent with PPT criterion. **(e)** The match rate of PPT criterion and prediction by tomographic predictor (255 features) when $p_{\min} = 0$.

When the number of neurons in the hidden layer becomes sufficiently large, the Bell-like predictor is capable of distinguishing the two classes of states with more than 99% in match rate, when $p_{\min} = 0.1$. In other words, the ensembles assumed to be separable or entangled can be classified reliably by our predictor at a very high accuracy.

To investigate further, we divide the data into three groups.

- **Group I:** The subclass of quantum states in ρ_{mix} in Eq. (26), in which the minimal eigenvalue, of the partial-transposed density matrix, is negative. These states are all entangled and all of them can be correctly labelled by PPT criterion.
- **Group II:** The complementary class of states of group I for ρ_{mix}, i.e., with non-negative eigenvalues. These states in principle can contain both entangled and separable states. There is currently no known method to identify the entanglement class of these states. They are most likely dominated by separable states, as p is very close to 0. However, we deliberately label all of them as entangled. We shall see that machine-learning method will suggest that most of them is separable, which is consistent with the prediction by PPT.
- **Group III:** The class of all fully-separable quantum states ρ_{sep}. These states are all separable, and all of them can be also correctly predicted by PPT criterion.

Our goal is to study the performance of the machine learning method in analyzing the internal structure of mixed quantum states. In the training phase, we labelled all states in group I and II with Green, and states in group III with Blue. Fig. 4**e** illustrates the result of the states with $p_{\min} = 0$, trained by the tomographic predictor. The length of bars represents the number of ensembles tested. For group I, which are definitely entangled, our tomographic predictor detects their entanglement with more than 99.6% (Ent. in figure) accuracy in terms of the match rate, which grows up to 99.9% (Sep. in figure) for the fully-separable states in group III. $\text{Bell}_{\text{ml}}(4, 80, 15)$ gives nearly the same result.

For group II, although the states were labeled as Green, tomographic predictor suggests that most of them are actually Blue (separable) states, which is consistent with the PPT criterion. This result demonstrates that the machine-learning method can potentially be used to identify if some states are labelled incorrectly.

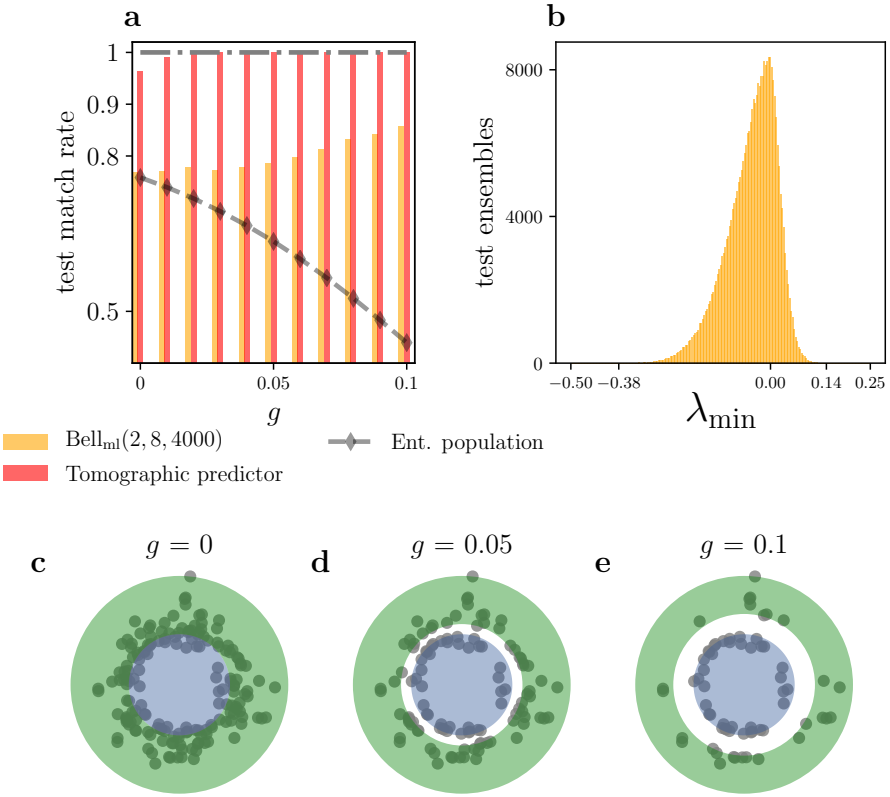


FIG. 5: **Bell-like & Tomographic predictor on all 2-qubit states. (a)** Match rate of machine learning predictors on the 2-qubit test ensembles that generated according to Harr distribution and then filtered out the entangled states that close to separable ones. The ratio (black dashed line) of entanglement ensembles to all states falls below 0.5 with $g \geq 0.09$. The match rate of tomographic predictor grows up to 99% when $g > 0.01$, and that of Bell-like predictor also grows with g increasing. **(b)** Distribution of $\lambda_{\min}(\rho_{AB}^{T_B})$ for test data. The distribution of training data with $g = 0$ is identical to but the size is 10 times larger than that of test data. **(c-e)** illustrate the state distribution. The states with $-g < \lambda_{\min} < 0$ are removed. The dots in the Green (Blue) area represent entangled (separable) states.

VII. GENERAL TWO-QUBIT STATES CLASSIFIED BY BELL-LIKE PREDICTORS

As we have claimed in the main text, tomography is necessary for general entanglement detection [9]. In this section, Bell-like predictor is implemented to be capable to classify *some* entangled states and all separable states, just like Bell inequality or other traditional methods in general. For this part, we generate a new training set of 2-qubit mixed states randomly and label them by the PPT criterion in the same way as the main text. The ensembles ρ are prepared by first generating a set of random matrices σ, where the real and imaginary parts of the elements $\sigma_{ij} = a_{ij} + ib_{ij}$ are generated by a Gaussian distribution with a zero mean and unit variance. The resulting density matrix is obtained by $\rho_{\mathrm{rand}} = \sigma\sigma^\dagger / \mathrm{tr}(\sigma\sigma^\dagger)$.

The performance of our machine-learning predictors heavily depends not only on the training set but also the distribution of the testing data. We find that many data points are localized near the boundary between entangled and separable states, which represents a challenge for us; machine learning performs not so well near the marginal cases. To regulate the anomaly, we introduce a gap in the boundary between the entangled and separable states.

Here the gap is defined for the entangled states where the absolute value of the minimum eigenvalue $\lambda_{\min}(\rho_{AB}^{T_B})$ of the partial transposed matrix $\rho_{AB}^{T_B}$ is larger than a given constant g, i.e.,

$$|\lambda_{\min}| > g\,. \tag{27}$$

The distribution of $\lambda_{\min}$ in our data set is given in Fig. 5**b**. We can see that the majority of states are weakly entangled, which imposes the challenge for our machine-learning predictors. The gap is used in both training and testing phase.

For this setting, we first present the results of $\mathrm{Bell}_{\mathrm{ml}}(2,8,4000)$ taking $\{\langle ab\rangle, \langle a'b'\rangle, \langle ab'\rangle, \langle a'b\rangle, \langle a\rangle, \langle b\rangle, \langle a'\rangle, \langle b'\rangle\}$ as features. Note that the measurement resources of $\mathrm{Bell}_{\mathrm{ml}}(2,8,4000)$ is same with $\mathrm{Bell}_{\mathrm{ml}}(2,4,4000)$ because the

TABLE I: features of machine learning predictors

Name	Features	Description
Tomographic Predictor	$\boldsymbol{O^1O^2}\cdots\boldsymbol{O^n}$, $\boldsymbol{O}\in\{I,\sigma_x,\sigma_y,\sigma_z\}$	Similar to n-qubit state tomography
$\mathrm{Bell}_{\mathrm{ml}}(n,3^n-1,x)$	$\boldsymbol{O^1O^2}\cdots\boldsymbol{O^n}$, $\boldsymbol{O}\in\{I,\hat{\boldsymbol{n}},\hat{\boldsymbol{n}}'\}$	x is neuron number in hidden layer 3^n-1 represents feature number
$\mathrm{Bell}_{\mathrm{ml}}(2,4,x)$	$\boldsymbol{ab}$, $\boldsymbol{ab'}$, $\boldsymbol{a'b}$, $\boldsymbol{a'b'}$	Similar to CHSH inequality
$\mathrm{Bell}_{\mathrm{ml}}(2,8,x)$	$\boldsymbol{ab}$, $\boldsymbol{ab'}$, $\boldsymbol{a'b}$, $\boldsymbol{a'b'}$, $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{a'}$, $\boldsymbol{b'}$	Same measurement resources with $\mathrm{Bell}_{\mathrm{ml}}(2,4,x)$
$\mathrm{Bell}_{\mathrm{ml}}(3,4,x)$	$\boldsymbol{abc}$, $\boldsymbol{a'b'c}$, $\boldsymbol{ab'c'}$, $\boldsymbol{a'bc'}$	Similar to Mermin inequality
$\mathrm{Bell}_{\mathrm{ml}}(3,8,x)$	$\boldsymbol{O^1O^2O^3}$, $\boldsymbol{O}\in\{\hat{\boldsymbol{n}},\hat{\boldsymbol{n}}'\}$	Svetlichny(Double Mermin) inequality
$\mathrm{Bell}_{\mathrm{ml}}(3,12,x)$	$\boldsymbol{abI}$, $\boldsymbol{ab'I}$, $\boldsymbol{a'bI}$, $\boldsymbol{a'b'I}$, $\boldsymbol{aIb}$, $\boldsymbol{aIb'}$, $\boldsymbol{a'Ib}$, $\boldsymbol{a'Ib'}$, $\boldsymbol{Iab}$, $\boldsymbol{Iab'}$, $\boldsymbol{Ia'b}$, $\boldsymbol{Ia'b'}$	Triple CHSH inequality

measurement outcomes of $\langle a\rangle, \langle b\rangle, \langle a'\rangle, \langle b'\rangle$ can be calculated directly without extra experimental resources. As shown in Fig. 5**a**, the match rate is about 75% when a gap is not introduced in the testing data. Unfortunately, the match rate cannot be considered as high, as the population of entangled states in the testing set is also about 75%; one can achieve the same performance by guessing all given states as entangled. This implies that the information involved in $\mathrm{Bell}_{\mathrm{ml}}(2,8,4000)$ features are not sufficient to detect entangled states in the ensemble. For a comparison, we consider using the tomographic predictor, which takes all 15 combinations of the two-qubit Pauli operators as an input. The result of match rate goes up to larger than 96%. This result implies that machine learning becomes more reliable when more information about the quantum states are available.

In addition, our Bell-like predictor gets improved if the entangled states near the boundary of the separable states are filtered out, as shown in Fig. 5**a**. Here we vary the gap from 0 to 0.1. The larger the gap, the better the performance of the Bell-like predictor. When the gap is about 0.08, the fraction of the entangled states is about the same as separable states, the match rate of the Bell-like predictor is about 85%.

VIII. DETAILS ABOUT OUR ARTIFICIAL NEURAL NETWORK (ANN) MODEL

If not special specified, all of our ANN model consists of 3 layers - input, hidden and output layer. Every layer contains a fixed-size block of units (neurons) which connected with units in the neighbor layers.

1. **input layer**. Here we denote the input layer as a vector $\vec{x}$. Our model encodes the observation results on the input layer, which includes some or all information of quantum system. Here the information - or more formally speaking - features, can be described by the expectation of observations (every single observation or observation products) . We assume every party observes and projects the system by 2 or 3 single Pauli-matrix, which can combine to the elements of Bell inequalities or standard state tomography. This is why the models are called Bell-like (2 observation per party) and tomographic (3 observation per party) predictor. TABLE 1 compares feature structure of our Machine learning predictors individually.

2. **hidden layer**. The hidden layer consists of neurons which denoted as $\vec{x}_1$.

$$\vec{x}_1 = \sigma_{RL}(W_1\vec{x} + \vec{w}_{10})\,, \tag{28}$$

where W_1, $\vec{w}_{10}$ are initialized uniformly and will be trained to decrease the loss function we will introduce soon. σ_{RL} is a nonlinear active function for every neuron. Here ReLu [10] are used as output function to the next layer. The definition of ReLu function σ_{RL}:

$$\sigma_{RL}([z_1, z_2, \cdots, z_{n_e}]^T) = [\max(z_1,0), \max(z_2,0), \cdots, \max(z_{n_e},0)]^T\,. \tag{29}$$

In our model, n_e represents the number of neurons in the hidden layer and $\vec{x}_1 = [z_1, z_2, \cdots, z_{n_e}]^T$.

3. **output layer**. This layer only contains 1 neuron for 2-class identification or m neurons for m-class identification, denoted as $\vec{x}_2$.

$$\vec{x}_2 = \sigma_S(W_2\vec{x}_1 + \vec{w}_{20})\,. \tag{30}$$

Here σ_S is sigmoid function $\sigma_S(z) = 1/(1+e^{-z})$ for 2-class identification or Softmax function for m-class identification.

$$\sigma_S([z_1, z_2, \cdots, z_m]^T) = [\sigma_S(z_1), \sigma_S(z_2), \cdots, \sigma_S(z_m)]^T\,, \tag{31}$$

$$\sigma_S(z_l) = \frac{e^{z_l}}{\sum_{j=1}^{m} e^{z_j}}, l = 0, 1, \cdots m \,. \tag{32}$$

Each neuron of output layer describes the probability (defined as c) of every class for given quantum state predicted by model.

In our work, we apply categorical cross entropy for multi-class classification (or binary cross entropy for binary classification) as loss function to quantify the difference between predictions and true categories (i.e. labels). Our models are trained by minimizing the loss function (the sum of Eq. 33 for all data, here $\hat{c}_j$ represents the probability of one state to be the j-th class predicted by machine, while c_j is that for label. c_j is either 1 or 0.) from the output layer by modifying the weights between neighbor layers.

$$H = -\sum_{j=1}^{m} c_j \log_2(\hat{c}_j) \,. \tag{33}$$

Now we just give an example to show the calculation of categorical cross entropy. For 4-type state classification task (Sec. V), separable states are labelled as $[1,0,0,0]^T$ while the other 3 types of entangled (biseparable) states are $[0,1,0,0]^T$, $[0,0,1,0]^T$ and $[0,0,0,1]^T$. If the model outputs $[0.9, 0.03, 0.03, 0.04]^T$, it actually means "the state is separable with probability = 90%." If the state is actually separable, the cross entropy is $-1 \times \log_2(0.9) - 0 \times \log_2(0.03) - 0 \times \log_2(0.03) - 0 \times \log_2(0.04)$. For 2-class identification, only one unit is applied in the output layer to output result.

ANN architecture can be easily constructed from readily-available tools such as Keras [11], which our model is implemented by. See Fig. 6 for more training details.

IX. KERNEL CODES OF GENERATING DATA FOR QUBIT SYSTEM

In our work, All the random operators and density matrices generated accroding to Haar measure, see more details below or in the website of QETLAB [5].

Our codes depend on Matlab to realize the randomness of matrices and wave functions (states). Here we just simplify and copy the kernel codes to show the method to generate these matrices and states. The size of generated matrix (state) = dim × dim (dim × 1). $1i = i$ is imaginary unit.

Random Unitary Matrix(return U)

```
gin = randn(dim) + 1i*randn(dim);
% Here gin is a random invertible matrix. (All the generated matrices are invertible.) Then apply QR decomposition of the Ginibre ensemble.
[Q,R] = qr(gin);
% compute U from the QR decomposition
R = sign(diag(R));
R(R==0) = 1;
% protect against potentially zero diagonal entries
U = bsxfun(@times,Q,R.');
% much faster than (but equivalent to) the naive U = Q*diag(R)
```

Random Density Matrix(return rho)

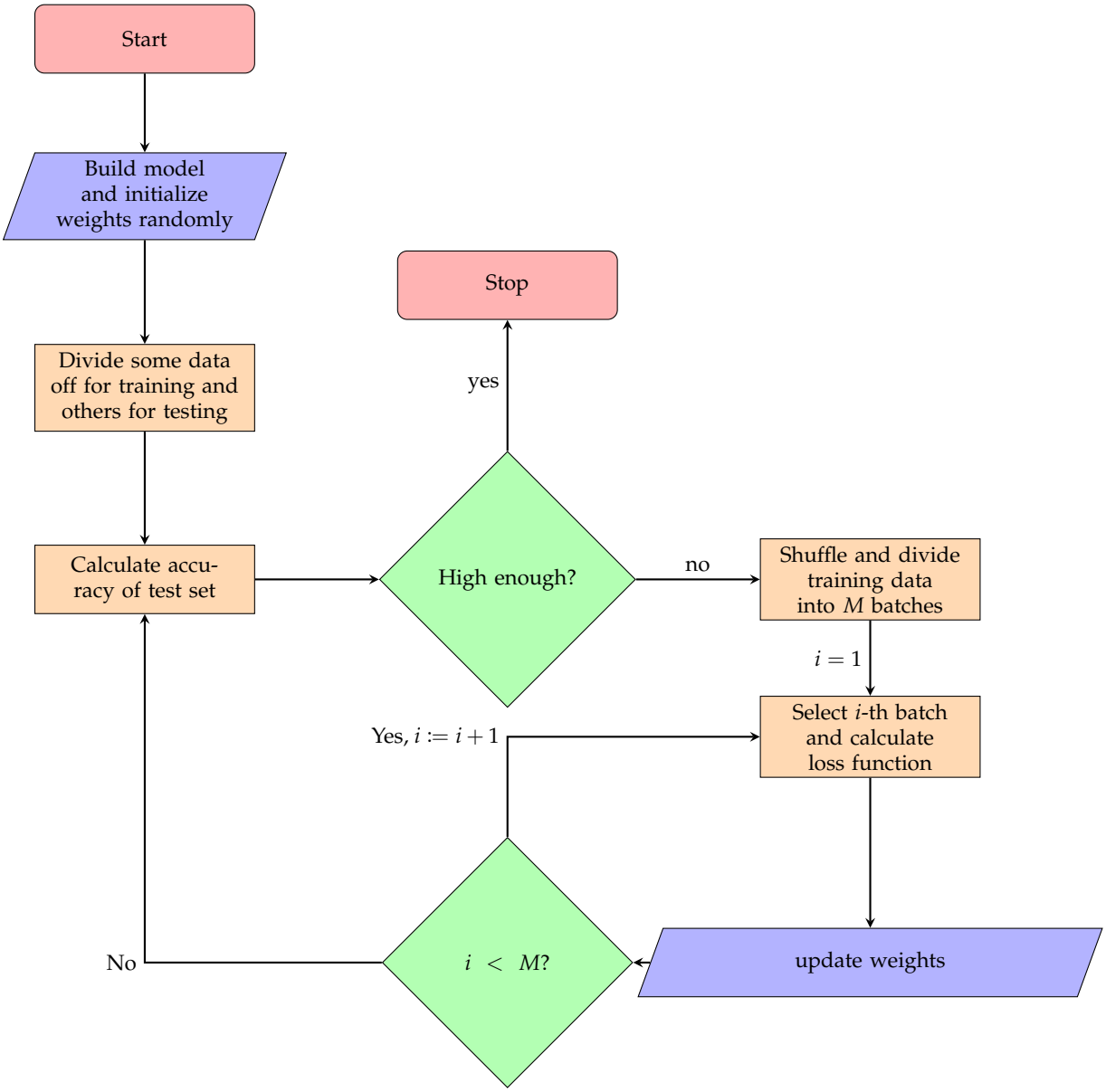
```
gin = randn(dim) + 1i*randn(dim);
rho = gin*gin';
% Here ' is dagger.
rho = rho/trace(rho);
```

Random State Vector(return v)

```
v = randn(dim,1) + 1i*randn(dim,1);
v = v/norm(v);
```

[1] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. Separability of mixed states: necessary and sufficient conditions. *Physics Letters A*, 223(1-2):1–8, 1996.

FIG. 6: Process of Machine Learning training



[2] J. S. Bell. Speakable and Unspeakable in Quantum Mechanics. *American Journal of Physics*, 57(6):567, 1989.

[3] Boris S Cirel'son. Quantum generalizations of bell's inequality. *Letters in Mathematical Physics*, 4(2):93–100, 1980.

[4] O Gühne, P Hyllus, D Bruß, A Ekert, M Lewenstein, C Macchiavello, and A Sanpera. Experimental detection of entanglement via witness operators and local measurements. *Journal of Modern Optics*, 50(6-7):1079–1102, 2003.

[5] Nathaniel Johnston. QETLAB: A MATLAB toolbox for quantum entanglement, version 0.9. http://qetlab.com, January 2016.

[6] Dafa Li, Xiangrong Li, Hongtao Huang, and Xinxin Li. Simple criteria for the slocc classification. *Physics Letters A*, 359(5):428–437, 2006.

[7] Otfried Gühne and Géza Tóth. Entanglement detection. *Physics Reports*, 474(1):1–75, 2009.

[8] Rob A Dunne and Norm A Campbell. On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In *Proc. 8th Aust. Conf. on the Neural Networks, Melbourne, 181*, volume 185, 1997.

[9] Dawei Lu, Tao Xin, Nengkun Yu, Zhengfeng Ji, Jianxin Chen, Guilu Long, Jonathan Baugh, Xinhua Peng, Bei Zeng, and Raymond Laflamme. Tomography is necessary for universal entanglement detection with single-copy observables. *Physical review letters*, 116(23):230501, 2016.

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. *Aistats*, 15:315–323, 2011.

[11] François Chollet et al. Keras. https://github.com/keras-team/keras, 2015.