Finding two-level systems in glasses through machine learning

Simone Ciarella,1,∗ Dmytro Khomenko,2,3,∗ Ludovic Berthier,4,5 Felix C. Mocanu,1 David R. Reichman,2 Camille Scalliet,6 and Francesco Zamponi1

1Laboratoire de Physique de l’École Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, 75005 Paris, France
2Department of Chemistry, Columbia University, 3000 Broadway, New York, NY 10027, USA
3Dipartimento di Fisica, Sapienza Università di Roma, P.le A. Moro 2, I-00185, Rome, Italy
4Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
5Laboratoire Charles Coulomb (L2C), Université de Montpellier, CNRS, 34095 Montpellier, France
6DAMTP, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom

(Dated: January 5, 2023)

Two-level systems (TLS) are rare quantum tunneling defects which govern the physics of glasses at very low temperature. Because of their extremely low density, it is very hard to directly identify them in computer simulations of model glasses. We introduce a machine learning approach to efficiently explore the potential energy landscape of glass models and identify two-level tunneling defects. We design an algorithm that is able to rapidly predict the quantum splitting between any two amorphous configurations produced by classical simulations. This in turn allows us to shift the computational effort towards the collection and identification of a larger number of TLS, rather than the useless characterization of non-tunneling defects which are much more abundant. Finally, we interpret our machine learning model to understand how TLS are identified and characterized, thus giving physical insight into the features responsible for their presence.

∗ These authors contributed equally. Email: simone.ciarella@ens.fr, dmytro.khomenko@uniroma1.it

arXiv:2212.05582v2 [cond-mat.dis-nn] 4 Jan 2023
I. INTRODUCTION

When a glass-forming liquid is cooled rapidly, its viscosity increases dramatically and it eventually transforms into an amorphous solid, called a glass, whose physical properties are profoundly different from those of ordered crystalline solids [1]. At even lower temperature, around 1K, the specific heat of a disordered solid is much larger than that of its crystalline counterpart as it scales linearly rather than cubically with temperature. Similarly, the temperature evolution of the thermal conductivity in glasses is quadratic, rather than cubic [2–11]. A theoretical framework rationalizing such anomalous behavior was provided by Anderson, Halperin and Varma [12] and by Phillips [13, 14]. They argued that the energy landscape of amorphous solids contains many nearly-degenerate minima, connected by localized motions of a few atoms, that can act as tunneling defects, called two-level systems (TLS). Understanding the microscopic origin of TLS and how to control their density and physical properties has gained significant interest, not only because they provide a large contribution to the specific heat and to the thermal conductivity, but also because their presence may impact the performance of certain quantum devices [15]. Unfortunately, understanding their microscopic nature and their almost ubiquitous presence in low-temperature glasses remains a challenge [16–20].

With the development of the swap Monte Carlo algorithm [21, 22] it has become possible to create computer glasses at unprecedentedly low temperatures. Combined with landscape exploration algorithms [23–32] this provides a method to investigate the nature of TLS in materials prepared under conditions comparable to experimental studies [33, 34]. These tools have enabled computational studies that have confirmed the experimental observation [7, 8, 10, 11, 35] that as the kinetic stability of a glass increases, the density of tunneling defects is strongly depleted [33, 36]. The direct detection of TLS revealed some of their microscopic features, namely that the participation ratio decreases in more stable glasses [33], and that TLS do not seem to be in a one-to-one correspondence with soft harmonic [33, 34, 37] or localized [34, 38] modes.

The main issue that limits the applicability of the direct landscape exploration method is that it remains computationally very expensive, and it is thus hard to construct the large library of TLS needed to enable a robust statistical analysis of their physical properties. After accumulating a large number of inherent structures (IS), one must run an expensive algorithm to find the relaxation pathway connecting pairs of IS to then determine if this pair forms a proper TLS (namely, has an energy splitting within thermal energy at 1K). Due to the large number of IS pairs detected, it is impossible to characterize all of them. In previous work, some ad hoc filtering rules were introduced [33, 34, 36] but the success rate of such filters is poor. After a significant exploration effort, it was possible to identify about 60 TLS starting from the identification of $\sim 10^8$ IS. It is then obvious that most of the computational effort has been wasted in the study of pairs that form defects which do not tunnel at low temperatures.

In this paper we show that it is possible to predict with enhanced accuracy whether a pair of inherent structures forms a TLS by using machine learning techniques. Recently, machine learning [39–48] has been shown to be extremely effective in using structural indicators to predict structural, dynamical or mechanical properties of glassy systems. In a similar vein, we use supervised learning to streamline the process of TLS identification. Our study has two goals: (i) develop a faster way to identify TLS compared to the standard approach outlined in Ref. [33] and described in Sec. II, in order to collect a statistically significant number of TLS; (ii) understand what are the structural and dynamical features characterizing TLS, and if/how they change with the preparation temperature. To address (i) we show that our machine learning model can be trained in a few minutes using a small quantity of data, after which the model is able to identify candidate TLS with high speed and accuracy. To address (ii) we show which static features are the most important for the model prediction and we show that states which are dynamically distant can nevertheless be TLS. We conclude by explaining how the ML model distinguishes TLS from non-TLS and how it is able to identify glasses prepared at different temperatures.

II. STANDARD APPROACH TO TLS IDENTIFICATION

The standard procedure [30, 31, 33] to identify TLS is sketched in Fig. 1. It consists of the following first steps which aim at identifying potential candidates for TLS:

1. Equilibrate the system at the preparation temperature $T_f$. Glasses with lower $T_f$ have a larger glass stability.

2. Run molecular dynamics to sample configurations along a dynamical trajectory at the exploration temperature $T < T_f$.

3. Perform energy minimization from the sampled configurations to produce a time series of energy minima, or inherent structures (IS).

4. Analyze the transition rates between pairs of IS, and select the pairs of IS that are explored consecutively.
Figure 1. Sketch of algorithms for TLS identification. The rough landscape of glassy models sketched in (a) limits the area that can be explored by each trajectory (black lines). In (b) a glass metabasin prepared in equilibrium at $T_f$ is sampled dynamically at $T < T_f$, leading to the identification of a large number of inherent states. The standard approach selects $\mathcal{O}(N_{IS})$ pairs of IS that are explored consecutively multiple times in the dynamics. The machine-learning approach (c) instead probes all $\mathcal{O}(N_{IS}^2)$ pairs and rapidly provides a robust prediction for their quantum splitting. The TLS candidates selected by the ML model are then analyzed via path-finding protocol (NEB calculation) and their quantum properties are computed from an effective one-dimensional Schrödinger equation.

Step 4 was necessary because it is computationally impossible to analyze all pairs of IS, as the number of pairs scales quadratically with the number of minima. The filter defined in step 4 was physically motivated by the fact that TLS tend to originate from IS that are not too distinct in order to have a reasonable tunneling probability. As such it is likely that those pairs of IS get explored one after the other during the exploration dynamics in step 2. Overall, given $N_{IS}$ inherent structures, this procedure selects for $\mathcal{O}(N_{IS})$ pairs to be analyzed.

However, many pairs of IS can be close but not sampled consecutively during the dynamics, owing to the complex structure of the potential energy landscape. Instead, our machine learning approach allows to consider all pairs of IS. As shown below, our approach is able to detect TLS which are otherwise excluded by the above step 4.

Once potential candidates are selected, the procedure continues as follows:

5. For each selected pair of IS, look for the minimum energy path and the classical barrier between them by running a minimum energy path finding algorithm, such as nudge elastic band (NEB) [49–51]. This provides the value of the potential along the minimum energy path between the pair $V(\xi)$, where $0 \leq \xi \leq 1$ is the reaction coordinate.

6. Select pairs whose energy profile $V(\xi)$ has the form of a double well (DW), i.e., exclude paths with multiple wells.

7. Solve the one-dimensional Schrödinger equation:

$$-rac{\hbar^2}{2md^2}\frac{\partial^2}{\partial \xi^2}\Psi(\xi) + V(\xi)\Psi(\xi) = E\Psi(\xi),$$  \hspace{1cm} (1)

where $\xi$ is a normalized distance along the reaction path $\xi = x/d$ and energy is normalized by a Lennard-Jones energy scale $\epsilon$, the effective mass $m$ and the distance $d$ are calculated as in Ref. [33]. We obtain the quantum splitting (QS) $E_{qs} = \mathcal{E}_2 - \mathcal{E}_1$ from the first two energy levels $\mathcal{E}_1$ and $\mathcal{E}_2$.

The quantum splitting is the most relevant parameter because when $E_{qs} \sim T$ the system can go from one state to the other via quantum tunneling [13], creating what is called a two-level system (TLS). In particular since we choose to report the data in units that correspond to Argon [33], a double well excitation will be an active TLS at $T = 1K$ when $E_{qs} < 0.0015\epsilon$, where $\epsilon$ sets the energy scale of the pair interactions in the simulated model.

Overall, since at low temperature the landscape exploration dynamics is slow, one would like to spend most of the computational time by doing steps 2-3 to construct a large library of pairs of IS. A first problem is that when the library of IS grows larger it takes a lot of time to perform steps 5-7. Moreover, the main bottleneck lies in the fact that most of the pairs that go through the full procedure are not found to be TLS in the end, and so this large computational time is effectively wasted.
III. MACHINE LEARNING APPROACH TO TLS IDENTIFICATION

As in any machine learning (ML) approach, we distinguish two phases: training and deployment. Our supervised training approach, detailed in the next section, takes just a few hours of training on a single CPU. It requires an initial dataset of $\mathcal{O}(10^4)$ full NEB calculations, whose collection is the most time consuming part of the training phase. Once training is complete, the ML model can be deployed to identify new TLS. Its workflow is similar to the standard one, with some major improvements. It proceeds with the following steps:

1. - 3. The first 3 steps are similar to the standard procedure to obtain a collection of inherent structures from a dynamical exploration.

4. Apply the ML model to all possible pairs of IS to predict which pairs form a DW potential.

5. Apply the ML model to predict the quantum splitting (QS) for all predicted DW and filter out the configurations that are not predicted to be TLS by the ML model.

6. - 8. Run NEB, select pairs that form DW potential and solve the one-dimensional Schrödinger equation for the TLS candidates only in order to obtain the exact value of the quantum splitting.

It is possible to use steps 4-5 as a single shot or as an iterative training approach (see Sec. IV D). In the Supplementary Information (SI) we provide details on how steps 1-3 are performed: glass preparation, exploration of the potential energy landscape via molecular dynamics simulations and minimization procedure, as well as NEB computation, see Sec. S1.

Noticeably, if the model is well-trained then the ML approach has two significant advantages over the standard approach. First, $\mathcal{O}(N_{IS}^2)$ pairs of IS are scanned to identify TLS, compared to a much smaller number $\mathcal{O}(N_{IS})$ in the standard procedure. Second, if a pair of IS passes step 5 and goes through the full procedure it is very likely to be a real TLS. As a consequence, by using the ML approach one can spend more time doing steps 2-3 to produce new IS, since fewer pairs pass step 5. At the same time, for any given number of IS, the ML approach can analyze all possible pairs and is therefore able to identify many more TLS, as we demonstrate below.

IV. MACHINE LEARNING MODEL

In Refs. [33, 34], the authors analyze a library of 14202, 23535 and 117370 pairs of inherent structures for a continuously polydisperse system of soft repulsive particles, equilibrated at reduced temperatures $T = 0.062, 0.07,$ and 0.092, respectively. The standard approach (Sec. II) leads to the identification of 61, 291 and 1008 TLS for the three temperatures, respectively. Notice that this approach uses pairs of IS that are selected by the dynamical information contained in the transition matrix between pairs of IS [33]. This was done to filter out all non double well potential. For all pairs in this small subset, the quantum splitting was then calculated.

Instead, the ML approach starts by independently evaluating the relevant information contained in each IS and constructs all possible combinations, even for pairs that are not dynamically connected. Following the steps discussed in Sec. III the model is then able to predict the quantum splitting of all the pairs, that were predicted to form a DW, very accurately. From a quantitative perspective, this means that the same trajectories now contain many more TLS candidates in the ML approach compared to the standard approach.

In this section we describe the flowchart of the model summarised in Fig. 2. We first discuss how we construct the dataset and extract the relevant features (Sec. IV A). We then explain the two main blocks consisting in a DW classifier followed by a quantum splitting predictor, both of which have similar model architecture and inputs. Finally we evaluate the performance of this model by showing its ability to accurately predict the quantum splitting. We conclude by introducing the iterative training technique that represents the optimal way to efficiently expand the library of TLS.

A. Dataset and features construction

The first step is the evaluation of a set of static quantities for all the available IS. This set consists of: energy, particle positions, averaged bond-orientational order parameters [52] determined via a Voronoi tessellation, from $q_2$ to $q_{12}$, and finally particle radii. The cost of this operation scales as the number of available states $N_{IS}$, but we use these quantities to calculate the features of $\sim N_{IS}^2$ pairs. A detailed analysis (see Sec. S3 in the SI) shows that the bond orientational parameters and the particle sizes are not very useful for the ML model. Since their calculation is slower than all the other features, we do not include them in the final version of the ML approach.
Flowchart of the machine learning approach. The first block represents the construction of the dataset by comparing all the pairs of inherent structures, focusing on the $M$ particles that displace the most. Then, required features are extracted to construct the input vector $X$. We then train a classifier to predict whether a pair of IS is a DW or not. The DW are finally processed using a multi-layer stacking strategy to predict the quantum splitting energy. Our pipeline analyses a given pair of IS in about $\sim 10^{-4}$ s.

To construct the input features for each pair of IS we combine the information of the two states evaluating the following:

- Energy splitting $\Delta E$: energy difference between the two IS.
- Displacements $\Delta \vec{r}_i$: displacement vector of particle $i$ between the two configurations.
- Total displacement $d$: total distance between the two IS defined as $d^2 = \sum_i |\Delta \vec{r}_i|^2$.
- Participation ratio $PR$: defined as $PR = (d^2)^2 / (\sum_i |\Delta \vec{r}_i|^4)$.
- Distance from the displacement center $|\vec{r}_0 - \vec{r}_i|$: we measure the average distance of particle $i$ from the center of displacement $\vec{r}_0$, identified as the average position of the particle that moves the most. This quantity identifies the typical size of the region of particles that rearrange.
- Transition matrix $T_{ij}$ (and $T_{ji}$): number of times that the exploration dynamics traverses IS$_j$ just after IS$_i$ (and vice versa).

The crucial step of the feature construction is that we can reduce the number of features by considering only the $M$ particles whose displacement is the largest between pairs of IS. We make this assumption because we expect that the low temperature dynamics is characterized by localised rearrangements involving only a small fraction of the particles [7, 8, 10, 11, 33, 35]. In Sec. S2 of the SI, we confirm this assumption by showing that the ML model achieves optimal performances even when $M$ is very small. So, the choice of $M \ll N$ makes the ML model computationally effective without any performance drop.

### B. Double well classifier

A necessary condition in order for a pair of IS to be a TLS is that the transition between the pair forms a double well (DW) potential. A DW is defined when the minimum energy path between the two IS resembles a quartic potential, as sketched in Fig. 1(c). The final goal of the ML model is to predict the quantum splitting of the pair to identify pairs with low values of the QS. The first obstacle in the identification of TLS is that DW represent only a small subgroup of all IS pairs. For instance in Ref. [33], only $\sim 0.5\%$ of all the IS pairs are DW at the lowest temperature. It is then mandatory to filter out pairs that are not likely to be a DW.

In the machine learning field there are usually many different models that can be trained to achieve similar performances, with complexity ranging from polynomial regression to deep neural networks. Here, we perform model ensembling and use ensembles both for DW classification and QS prediction. Model ensembling consists in averaging the output of different ML models to achieve better predictions compared to each of them separately. We do so using the publicly available AutoGluon library [53]. In this approach, we train in a few minutes a single-stack ensemble that is able to classify DW with $> 95\%$ accuracy. In Sec. S2 of the SI we justify this choice of ML model and provide details on performances and hyperparameters. Overall, since the DW classifier is accurate and rapid, we use it to filter out the pairs that do not require the attention of the QS predictor because they cannot be TLS anyway.
Figure 3. (a)-(c) Quantum splitting and (d)-(f) energy barrier predicted by the ML model compared to the exact value, reported using a double logarithmic scale. We report predictions for IS pairs that the ML model has not seen during the training. (a) and (d) correspond to \( T_f = 0.062 \), with training for 7000 samples and information of the \( M = 3 \) particles with largest displacements. (b) and (e) correspond to \( T_f = 0.07 \), 10000 samples and \( M = 3 \). (c) and (f) correspond to \( T_f = 0.092 \), 30000 samples and \( M = 3 \). All models have been trained for \( \sim 10 \) hours of single CPU time.

C. Quantum splitting predictor

We want to predict the quantum splitting of a pair of IS for which the features discussed in Sec. IV A have been computed. We need this prediction to be very precise, because we know that a pair can be considered a TLS when \( E_{qs} < 0.0015 \epsilon \), but \( E_{qs} \) can vary significantly so errors may be large. In the SI (see Sec. S2) we show that models such as deep neural networks and regression are not stable or powerful enough to achieve satisfying results. We thus perform model ensembling by using the AutoGluon library [53]. This achieves superior performances while also allowing us to compare and rank single models. We perform two types of model ensembling: (i) stacking: different layers of models are applied in series creating a stack (schematized in Fig. 2), and (ii) bagging: we divide the data in subsets that we use to train multiple instances of the same models and we later combine all their predictions. As shown in the SI, the best results are obtained while using ensembles of gradient boosting methods (in particular CatBoost [54]), which have proven to be the optimal choice in similar semi-empirical quantum-mechanical calculations [55]. Overall, the ensemble structure of our final model relieves us from hyperparameter optimization and makes the results more stable.

In order to train the model we first collect a set of \( E_{qs} \) examples. The size of this training set is discussed in the SI (see Fig. S2) where we find that the minimum number is around \( 10^4 \). We can use some of the data already collected in previous work in Ref. [33] for the training. Moreover, since we are interested in estimating with more precision the lowest values of \( E_{qs} \) we train the model to minimize the following loss function

\[
L = \frac{\sum_{i=1}^{n} w_i (E_{qs, \text{true}} - E_{qs, \text{predicted}})^2}{n \sum_{i=1}^{n} w_i},
\]

which is a weighted mean-squared error. The weights correspond to \( w_i = 1/E_{qs, \text{true}} \) in order to give more importance to low \( E_{qs} \) values. We thus train our model to provide a very accurate prediction of the value \( E_{qs} \) for any given pair. Once the model is trained it takes only \( \sim 10^{-4} \)s to predict the QS of a new pair (compared to 1 minute to run the standard procedure). If we predict a value \( E_{qs} < 0.0015 \epsilon \), then we have identified a TLS much faster.

To showcase the performance of the ML model we report in Fig. 3(a)-(c) the exact quantum splitting calculated from the NEB procedure, compared with the value predicted by the model. We have trained three independent models to work at the three different temperatures. As explained before (Sec. IV A), the model needs the information about only the \( M \ll N \) particles that are displaced the most to achieve the excellent precision demonstrated in Fig. 3.
the Fig. S2 (SI) we find that the optimal value is $M = 3$, confirming that the participation ratio in TLS is quite low, because only three particles are needed for the model to identify TLS. Furthermore, the models have been trained using the smallest number of samples, randomly selected from all the IS pairs available, that allows the model to reach its top performance. We have also performed an analysis of the optimal training time. Details on these points are provided in Sec. S2 of the SI. The performances presented in Fig. 3 are achieved by training the model for ~ 10 hours of single CPU time, but we also show in the SI that it is possible to already achieve > 90% of this performance by training the ensemble for only 10 minutes.

The ML approach that we have just introduced is also easily generalizable to target any state-to-state transition, like excitations and higher energy effects. Here we modified the quantum splitting predictor to instead predict the classical energy barrier between two IS states. If the minimal energy path between two IS forms a DW, we define the classical energy barrier as the maximum value of the energy along this path. In Fig. 3(d)-(f) we report the value of the energy barrier predicted by the ML model (y-axis) compared to the exact value calculated from the NEB procedure (x-axis). The hyperparameters and the features are the ones used for the quantum splitting predictor. Such a high performance demonstrates that our ML approach can predict other types of transitions between states.

### D. Iterative training procedure

We finally introduce an approach to optimally employ our ML model to process new data: the iterative training procedure. In previous sections and in Fig. 3 we trained the model once using a subset of the already available data. This is a natural way to proceed when the goal is to process new data that are very similar to the training set, and the training set is itself large enough. However, since the goal of the proposed ML model is to ultimately drive the landscape exploration and collect new samples, the single-training approach may encounter two types of problems. First, at the beginning there may be not enough data and second the findings of the model do not provide any additional feedback.

To solve both problems we introduce the iterative training procedure. The idea of iterative training is to use the predictive power of ML to create and expand its own training set, consequently enhancing its performance by iteratively retraining over the new data. Details on the method and parameters are discussed in Sec. S4 of the SI. In practice, we start from a training set of $K_0 \sim 10^3 - 10^4$ randomly selected pairs to have an initial idea of the relation between input and output. We then use the ML approach outlined in Fig. 2 to predict the $K_i = 500$ pairs with the lowest QS. For these TLS candidates, we perform the full procedure to calculate the true QS and determine whether the pair is a DW or a TLS. In Fig. S6 (SI), we report the result of this procedure when we process a new set of trajectories from the same polydisperse soft sphere potential as in Ref. [33]. In general the first few iterations of iterative training have a poor performance. In fact we find that > 70% of the first $K_i$ pairs are actually non-DW. After collecting additional $K_i$ measurements, we retrain the model. We report in Tab. I the average time for each step of the ML procedure. The retraining can be done in ~ 10 min, after which the model is able to predict the next $K_i$ pairs with lowest QS. Overall, to process $N_{\text{ISpairs}}$ we estimate that the computational time of the iterative approach is $t_i = \left[ K_0 \cdot 10^2 + N_{\text{iter}} (K_i \cdot 10^2 + 10^3 + N_{\text{ISpairs}} \cdot 10^5) \right] s$. If $N_{\text{ISpairs}} > 10^9$ it is possible to significantly reduce $t_i$ by permanently discarding the worst pairs, but this is not needed here. We iterate this procedure $N_{\text{iter}}$ times, until the last batch of $K_i$ candidates contains less than 1% of the total number of TLS. We believe that continuing this iterative procedure would lead to the identification of even more TLS/DW, but this is out of the scope of this paper.

| Procedure                              | Time (s)  |
|----------------------------------------|-----------|
| NEB+Schrödinger for a IS pair          | $10^2$    |
| Predict if a IS pair is a DW            | $10^{-5}$–$10^{-4}$ |
| Predict $E_{qs}$ for a IS pair          | $10^{-5}$–$10^{-4}$ |
| Train DW classifier                     | $10^{-4}$–$10^{-3}$ |
| Train QS predictor                      | $10^{-5}$–$10^{-4}$ |

Table I. Computational time needed to perform our ML approach, on a standard laptop.

### V. RESULTS AND DISCUSSION

We now use ML in order to speed up the TLS search. This highly efficient method allows us to collect a library of TLS of unprecedented size, generated from numerical simulations with the same interaction potential as in Ref. [33].

First of all, we reprocess the data produced to obtain the results presented in Ref. [33] with our new ML method, obtaining new information about the connection between TLS and dynamics. Next, we perform ML-guided exploration

```latex
\begin{table}
\centering
\begin{tabular}{|c|c|}
\hline
Procedure                              & Time (s)  \\
\hline
NEB+Schrödinger for a IS pair          & $10^2$    \\
Predict if a IS pair is a DW            & $10^{-5}$–$10^{-4}$ \\
Predict $E_{qs}$ for a IS pair          & $10^{-5}$–$10^{-4}$ \\
Train DW classifier                     & $10^{-4}$–$10^{-3}$ \\
Train QS predictor                      & $10^{-5}$–$10^{-4}$ \\
\hline
\end{tabular}
\caption{Computational time needed to perform our ML approach, on a standard laptop.}
\end{table}
```
to collect as many TLS as possible. This sizable library of TLS allows us to perform for the first time a detailed statistical analysis of TLS and compare their distribution to the distribution of double wells. We perform this analysis for glasses of three different stabilities. Finally, we discuss the microscopic features of TLS not only by looking at their statistics, but also by analyzing what the ML model has learned, and how it expresses its prediction.

A. Capturing elusive TLS with machine learning

Prior to this paper, it was not possible to evaluate all the IS pairs collected in Ref. [33]. For this reason the authors discarded \( a \text{ priori} \) all pairs where the number of forward and backward jumps between IS is such that \( \min(T_{ij}, T_{ji}) < 4 \), based on the assumption that high transition rates during the dynamic landscape exploration is a good indicator that two IS form a DW. This reduced the number of pairs to 14202, 21109 and 117339 for glasses prepared at \( T_f = 0.062 \), 0.07 and 0.092 respectively. In order to have comparable data at the three temperatures, for \( T_f = 0.092 \) we only consider a subset of glasses corresponding to 30920 IS pairs.

The results of the TLS search are summarized in the red columns of Tab. II. Overall the standard procedure reaches a rate of TLS found over calculations performed of \( 4 \cdot 10^{-3}, 13 \cdot 10^{-3}, \) and \( 8 \cdot 10^{-3} \) for the three temperatures, respectively. We compare this with our iterative training procedure applied on the same data, whose results are reported in the green columns of Tab. II. We immediately notice two major improvements: the overall number of TLS that we find from the same data set is more than twice larger and the ratio of TLS per calculation is more than 15 times larger, corresponding to \( 62 \cdot 10^{-3}, 211 \cdot 10^{-3}, \) and \( 194 \cdot 10^{-3} \) for the same three temperatures.

We conclude that the iterative ML approach is much more efficient than the standard procedure, and also that TLS do not necessarily have a large dynamical transition rate, since the dynamical-filtering approach misses more than half of them.

B. Differences between DW and TLS

With our ML-driven exploration of the energy landscape we can restrict the numerical effort to DW and favorable TLS candidates, while processing a larger number and/or longer exploration trajectories. This allows us to consider a larger set of independent glasses of the same type as those treated in Ref. [33], which is particularly relevant for ultrastable glasses generated at the lowest temperature \( T_f = 0.062 \). While in Ref. [33] the collection of 61 TLS required \( > 14000 \) NEB calculations, we are able to identify 864 TLS running 11 iterations of iterative training using only a total of 5500 NEB calculations in addition to the \( \sim 6000 \) used for pretraining. In the next section we analyze these results to discuss the nature of TLS.

The database of glasses that we analyze with iterative training contains 5 times more minima than in Ref. [33], but we are able to find 15 times more TLS running around half of the NEB calculations. Furthermore, in Fig. 4(a) we have color-coded the confusion matrix and we highlight that we can capture almost all of the TLS if we consider

| \( T_f \)   | \# TLS | \# NEBs | \# TLS | \# NEBs |
|------------|--------|---------|--------|---------|
| 0.062 (Ar) | 61     | 14202   | 156    | 2500    |
| 0.062 (NiP)| 28     | 14202   | 59     | 2000    |
| 0.07 (Ar)  | 291    | 21109   | 1057   | 5000    |
| 0.07 (NiP) | 46     | 21109   | 152    | 4000    |
| 0.092 (Ar) | 245    | 30920   | 776    | 4000    |
| 0.092 (NiP)| 28     | 30920   | 123    | 6000    |

* This number does not include the data that we use to pre-train the model.

Table II. Analysis of data collected in Ref. [33]. We compare the standard procedure with our ML approach using iterative training. We report results for different glass stabilities, decreasing from top to bottom, using Argon units (Ar) [28] and NiP metallic glass parameters (NiP) [25]. The standard procedure finds less than half of the TLS and is computationally much more expensive.
Figure 4. Results of ML driven exploration, leading to a library of TLS of unprecedented size. (a) We compare the model predictions to the calculated values at the end of our iterative training, at \( T_f = 0.062 \). We color coded in the background the confusion matrix. The black dashed horizontal lines report the percentage of TLS that are predicted below that value of quantum splitting, showing that more than 95% of the TLS are within twice the TLS threshold of 0.0015. (b) Cumulative distribution of energy splitting \( n(E_{qs}) \) divided by \( E_{qs} \). (c) Histograms of the number of TLS and DW per glass, at the three preparation temperatures. In total we have considered 237, 30, 5 glasses equilibrated at \( T_f = 0.062, 0.07, \) and 0.092 respectively.

only the pairs that are predicted to be within twice the quantum splitting threshold of TLS. In Fig. 4(b) we report the saturation plateau of the cumulative density of TLS quantum splitting \( n(E_{qs}) \), which scales as \( n(E_{qs}) \sim n_0 E_{qs} \) at low \( E_{qs} \). The ML approach allows us to collect significantly better statistics compared to Ref. [33], confirming that the TLS density \( n_0 \) decreases by several orders of magnitude from hyperquenched to ultrastable glasses. Lastly, in Fig. 4(c) we report the histograms of the number of TLS and DW per glass at the three temperatures. We see that when the glasses are ultrastable \( (T_f = 0.062) \) most of the glasses have very few TLS. Conversely, poorly annealed \( (T_f = 0.092) \) glasses show a very unbalanced distribution, with few glasses that contains most of the DW and TLS.

C. Interpretation of the ML model

It is possible to use the ML model to gain information about the distinctive structure of TLS. First, the present and previous works [7, 8, 10, 11, 33, 35, 36] find that the density of TLS decreases upon increasing glass stability, which in our simulations is controlled by the preparation temperature. Thus, one may also expect temperature-dependent TLS features. In the SI we show that when the ML model is trained at \( T_{\text{train}} \) and deployed at \( T_{\text{prediction}} \neq T_{\text{train}} \) there is only a minor performance drop and the model is able to perform reasonably well. This implies that the model captures distinctive signatures of TLS that do not depend strongly on the preparation temperature. Yet, we also show in the SI that it is very easy to train another ML model to predict the temperature itself and eventually add it to the pipeline.

Overall, the ML model is not only able to capture the different microscopic features of TLS, but it can also suggest what is the specific influence of each feature. To interpret this information we calculate their Shapley values [56] and we report them in Fig. 5. The features are ranked from the most important (top) to the less important (bottom) reporting the impact that they have on the model output (SHAP value), so that a positive SHAP value predicts on average a high value of the QS. We see that the most important feature is the classical energy splitting \( \Delta E \) and that a large splitting (red) corresponds to a large QS. The second most important feature is the largest single particle displacement \( \Delta r_0 \), which has to be larger than a threshold corresponding to 0.3\( \sigma \) in order to predict a low QS. The total displacement \( d \) is the third most important and shows a similar effect. All the remaining features have a less clear and much smaller effect on the model prediction and they only collaborate collectively to the final QS prediction.

In the SI we show that it is possible to obtain very good performance even when removing some of the features with the largest Shapley values, which means that the ML interpretation is not unique.

According to this Shapley analysis we explain the ML prediction in the following way: the energy difference \( \Delta E \) between two IS is the main predictor for the quantum splitting, and it has to be small for TLS. Then, the largest particle displacement \( \Delta r_0 \) is necessary to understand if the two IS are similar and what is their stability (we show in the SI that \( \Delta r_0 \) is the most important feature to identify the glass stability). Then the total displacement \( d \) complements this information and gives local information about the displacements of the other particles. Lastly, all the other inputs
Figure 5. Importance of the different features for the ML model for the prediction of the quantum splitting, evaluated from the Shapley values [56] at $T_f = 0.062$. The features are ranked from the most important (top) to the least important, where the impact that they have on the model output corresponds to their SHAP value. Each point corresponds to a single IS pair. The color coding shows the impact from high (red) to low (green) values of each specific input.

provide fine tuning to refine the final prediction and are discussed in more detail in the SI. Interestingly, in the SI we also show that even without the two most important features, the ML approach can still reasonably reveal TLS candidates.

D. Microscopic features of TLS

We have shown that by following a ML-driven approach it is possible to collect a significant library of TLS for any preparation temperature. However it may be useful to discuss alternative strategies to rapidly identify TLS. In general, since TLS are extremely rare objects [7, 8, 10, 11, 35] a filtering rule is necessary in order to reduce the number of possible candidates. In particular, Ref. [33, 34, 36] proposed to use the transition matrix to exclude pairs that do not get explored consecutively. This is based on the assumption that DW (and consequently TLS) correspond to IS pairs that are close to each other and therefore during an exploration run should be detected successively, hence have a non zero transition rates.

Instead, we prove in Fig. 6 that a filter based on dynamical information only is a poor predictor. In Fig. 6(a) we report the distribution of dynamical transitions between two inherent structures $i$ and $j$ for TLS, DW and all the pairs, measured at $T_f = 0.062$. While the slowly decaying tail of TLS and DW suggests that they often exhibit a large transition rate, actually most of the TLS and DW are characterized by no transition at all between them. Our interpretation is that even though the transition towards the other side of the double well is favourable, the landscape has such a large dimensionality ($3N$) that even very favorable transitions may never take place in a finite exploration time. This issue can become more severe when the trajectories are shorter, for example if the exploration is performed in parallel. We confirmed this observation by the results that are already reported in Tab. II, where we have used our iterative training approach to re-analyze the data of Ref. [33], including pairs with no transition, thus finding many more TLS. We conclude that even though the transition rates are the most important single characteristics for DW prediction (see Fig. 5), a filter based solely on them is still missing a lot of interesting pairs and therefore is not the most efficient.

In Fig. 6(b) we focus on the distribution of classical splitting $\Delta E$. When $\Delta E$ is large we rarely see DW and TLS (red region). On the other hand, there are many pairs that show a very small $\Delta E$ value (yellow region), but they are not more likely to be TLS. Ultimately we find a ‘sweet spot’ (green region), where TLS are more frequent. The ML model also captures this feature, as we can see from the SHAP parameter of $\Delta E$ in Fig. 5. The next most important feature according to the ML model is the largest particle displacement $\Delta \vec{r}_0$, reported in Fig. 6(c). When it is larger than $\sim 0.8\sigma$ we rarely find TLS and DW, but we do not find them also when $\Delta \vec{r}_0 < 0.3\sigma$. The SHAP parameter in Fig. 5 confirms that the ML model has discovered this feature. Finally in Fig. 6(d) we report the total displacement $d$. If $d > 0.9\sigma$ the pair is so different that it is not likely to be a TLS or DW, while this probability increases for smaller $d$.

Overall, if it is necessary to identify TLS with a ‘quick and dirty’ method, we propose to use a transition matrix to filter DW from non-DW, and then select a sweet spot for the energy splitting and the displacement for selecting optimal TLS candidates.
Figure 6. Microscopic features of TLS and DW for ultrastable glasses at $T_f = 0.062$. We report the probability distribution functions of: (a) number of transitions between the two inherent structures $T_{ij}$ and $T_{ji}$, (b) classical energy splitting $\Delta E$, (c) largest particle displacement $\Delta r_0$ and (d) total displacement $d$. We color coded in red the regions of parameters where we do not expect to find TLS, which instead concentrate in the green regions. These green regions could serve as an alternative to rapidly identify TLS.

VI. CONCLUSION

In this paper we have introduced a machine-learning approach to explore complex energy landscapes, with the goal of efficiently locating double wells (DW) and two-level systems (TLS). We demonstrate that it is possible to use ML to rapidly estimate the quantum splitting of a pair of inherent structures (IS) and accurately predict if a DW is a TLS or not. We also show that our ML approach can be used to predict very accurately the energy barrier between pairs of IS. Overall, this approach allows us to collect a TLS library of unprecedented size that would be impossible to obtain without the support of ML.

The ML model uses as input information calculated from a pair of inherent structures. After just a few minutes of supervised training it is able to infer with high accuracy the quantum splitting of any new pair of inherent structures. We establish that the ML model that we develop using the Autogluon library is fast and precise. Its efficiency allows us to introduce an iterative training procedure, where we perform a small batch of prediction and then retrain the model.

After performing statistical analysis over the unprecedented number of TLS collected with our method, we have discovered that many DW and TLS are not consecutively explored during the dynamics. We then reanalyzed the data of Ref. [33] finding that more than half of the TLS had been missed, because the analysis of Ref. [33] was based on this dynamic assumption. Our ML approach not only finds more than twice the number of TLS from the same data, but it also requires significantly fewer calculations. Overall we conclude that ML significantly improves the existing approaches. It also shows that if it is not possible to use the ML procedure that we outline, an effective ‘quick and dirty’ way to predict TLS should be: a) use $T_{ij}, T_{ji}$ for predicting DW; b) for predicted DW use energy splitting between two IS for predicting TLS.

We also discuss the microscopic nature of DW and TLS. We perform a Shapley analysis to dissect the ML model and understand what it learns, and we compare this with the extended statistics of TLS that we are able to collect. We find that the quantum splitting is mostly related to the classical energy splitting and the displacements of the particles. Overall, the Shapley analysis suggests that TLS are characterized by one particle that displaces between 0.3 and 0.9 of its size, while the total displacement and the energy difference between the two states remains small. The local structure around the particle is not as important, nor is the number of times we actually see this transition during the exploration dynamics.

Lastly we investigate the effect that glass stability (equivalent to the preparation temperature in our simulations) has on double wells and TLS. The ML model learns that at higher temperatures all the pairs are characterized by more collective rearrangements, but TLS are similar for any preparation temperature.

Ultimately, since our ML approach is extremely efficient in exploring the energy landscape and is easy to generalize to target any type of state-to-state transition (as we show for the energy barriers), we hope that our method will be used in the future to analyze not only TLS, but also many other examples of phenomena related to specific transitions between states in complex classical and quantum settings.
ACKNOWLEDGMENTS

We thank E. Flenner, G. Folena, M. Ozawa, J. Sethna, S. Elliott, G. Ruocco and W. Schirmacher for useful discussions. This work was granted access to the HPC resources of MesoPSL financed by the Region Ile de France and the project Equip@Meso (reference ANR-10-EQPX-29-01) of the programme Investissements d’Avenir supervised by the Agence Nationale pour la Recherche. This project received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program, Grant No. 723955 – GlassUniversality and the project Equip@Meso (reference ANR-10-EQPX-29-01) of the programme Investissements d’Avenir supervised by the Leverhulme Trust (VP1-2019-029, LB). CS acknowledges support from the Herchel Smith Fund and Sidney Sussex College, University of Cambridge.

[1] P. G. Debenedetti and F. H. Stillinger, Nature 410, 259 (2001).
[2] R. C. Zeller and R. O. Pohl, Physical Review B 4, 2029 (1971).
[3] M. T. Loponen, R. C. Dynes, V. Narayanamurti, and J. P. Garno, Physical Review B 25, 1161 (1982).
[4] J. F. Berret and M. Meißner, Zeitschrift für Physik B Condensed Matter 70, 65 (1988).
[5] J. M. Leveritt, G. Ruyters, C. Schötz, M. Bazarbashan, P. Fasll, M. von Schickfus, A. Fleischmann, and C. Enss, EPL (Europhysics Letters) 104, 57006 (2013).
[6] D. R. Queen, X. Liu, J. Karel, T. H. Metcalf, and F. Hellman, Physical Review Letters 110, 135901 (2013).
[7] T. P.-C. neda, R. J. Jiménez-Riobóo, and M. A. Ramos, Physical Review Letters 112, 165901 (2014).
[8] T. P.-C. neda, C. Rodríguez-Timoc, J. Rodríguez-Viejo, and M. A. Ramos, Proceedings of the National Academy of Sciences 111, 11275 LP (2014).
[9] X. Liu, D. R. Queen, T. H. Metcalf, J. E. Karel, and F. Hellman, Physical Review Letters 113, 25503 (2014).
[10] D. R. Queen, X. Liu, J. Karel, H. C. Jacks, T. H. Metcalf, and F. Hellman, Journal of Non-Crystalline Solids 426, 19 (2015).
[11] P. w. Anderson, B. I. Halperin, and c. M Varma, The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics 25, 1 (1972).
[12] W. A. Phillips, Journal of Low Temperature Physics 7, 351 (1972).
[13] W. A. Phillips, Reports on Progress in Physics 50, 1657 (1987).
[14] J. M. Martinis, K. B. Cooper, R. McDermott, M. Steffen, M. Ansmann, K. D. Osborn, K. Cicak, S. Oh, D. P. Pappas, R. W. Simmonds, and C. C. Yu, Physical Review Letters 95, 210503 (2005).
[15] A. J. Leggett and D. C. Vural, The Journal of Physical Chemistry B 117, 12966 (2013).
[16] D. Zhou, Journal of Physics: Condensed Matter 32, 55704 (2019).
[17] H. M. Carruzzo and C. C. Yu, Physical Review Letters 124, 75902 (2020).
[18] C. Artiaco, F. Balducci, and A. Scardicchio, Physical Review B 103, 214205 (2021).
[19] W. Ji, arXiv preprint arXiv:2112.10105 (2021).
[20] A. Ninarello, L. Berthier, and D. Coslovich, Physical Review X 7, 021039 (2017).
[21] L. Berthier, E. Flenner, C. J. Fullerton, C. Scalliet, and M. Singh, Journal of Statistical Mechanics 2019, 64004 (2019).
[22] F. H. Stillinger and T. A. Weber, Physical Review A 25, 978 (1982).
[23] T. A. Weber and F. H. Stillinger, Physical Review B 32, 5402 (1985).
[24] A. Heuer and R. J. Silbey, Phys. Rev. Lett. 70, 3911 (1993).
[25] D. Dab, A. Heuer, and R. J. Silbey, Journal of Luminescence 64, 95 (1995).
[26] A. Heuer and R. J. Silbey, Physical Review B 53, 609 (1996).
[27] F. Demichelis, G. Viliani, and G. Ruocco, PhysChemComm 2, 20 (1999).
[28] J. Reinsch and A. Heuer, Physical Review B 70, 64201 (2004).
[29] J. Reinsch and A. Heuer, Physical Review Letters 95, 155502 (2005).
[30] A. Heuer, Journal of Physics: Condensed Matter 20, 373101 (2008).
[31] T. Damart and D. Rodney, Physical Review B 97, 14201 (2018).
[32] D. Khomenko, C. Scalliet, L. Berthier, D. R. Reichman, and F. Zamponi, Physical Review Letters 124, 225901 (2020).
[33] D. Khomenko, D. R. Reichman, and F. Zamponi, Physical Review Materials 5, 55602 (2021).
[34] M. D. Ediger, Proceedings of the National Academy of Sciences 111, 11232 LP (2014).
[35] F. C. Mocanu, L. Berthier, S. Ciarella, D. Khomenko, D. R. Reichman, C. Scalliet, and F. Zamponi, arXiv preprint arXiv:2209.09579 (2022).
[36] L. Wang, A. Ninarello, P. Guan, L. Berthier, G. Szamel, and E. Flenner, Nature Communications 10, 26 (2019).
[37] B. B. Laird and H. R. Schober, Physical Review Letters 66, 636 (1991).
[38] S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras, and A. J. Liu, Nature Physics 12, 469 (2016).
[39] D. M. Sussman, M. Paoluzzi, M. C. Marchetti, and M. L. Manning, EPL (Europhysics Letters) 121, 36001 (2018).
[40] E. Boattini, M. Ram, F. Smallenburg, and L. Filion, Molecular Physics 116, 3066 (2018).
[41] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E. D. Cubuk, S. S. Schoenholz, A. Obika, A. W. R. Nelson, T. Back,
D. Hassabis, and P. Kohli, Nature Physics 16, 448 (2020).
[43] J. Paret, R. L. Jack, and D. Coslovich, The Journal of Chemical Physics 152, 144502 (2020).
[44] D. Richard, M. Ozawa, S. Patinet, E. Stanifer, B. Shang, S. A. Ridout, B. Xu, G. Zhang, P. K. Morse, J.-L. Barrat, L. Berthier, M. L. Falk, P. Guan, A. J. Liu, K. Martens, S. Sastry, D. Vandembroucq, E. Lerner, and M. L. Manning, Phys. Rev. Materials 4, 113609 (2020).
[45] E. Boattini, F. Smallenburg, and L. Filion, Physical Review Letters 127, 88007 (2021).
[46] S. A. Ridout, J. W. Rocks, and A. J. Liu, Proceedings of the National Academy of Sciences 119, e2119006119 (2022).
[47] I. Tah, S. A. Ridout, and A. J. Liu, The Journal of Chemical Physics 157, 124501 (2022).
[48] G. Jung, G. Biroli, and L. Berthier, arXiv preprint arXiv:2210.16623 (2022).
[49] G. Henkelman, B. P. Uberuaga, and H. Jonsson, The Journal of Chemical Physics 113, 9901 (2000).
[50] W. E, W. Ren, and E. Vanden-Eijnden, Physical Review B 66, 52301 (2002).
[51] J. Guéoné, W. G. Nöhring, A. Vaid, F. Houllé, Z. Xie, A. Prakash, and E. Bitzek, Computational Materials Science 175, 109584 (2020).
[52] W. Lechner and C. Dellago, The Journal of Chemical Physics 129, 114707 (2008).
[53] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, arXiv preprint arXiv:2003.06505 (2020).
[54] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, Advances in neural information processing systems 31 (2018).
[55] X. García-Andrade, P. G. Tahoces, J. Pérez-Ríos, and E. M. Núñez, arXiv preprint arXiv:2208.02289 (2022).
[56] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, Nature Machine Intelligence 2, 56 (2020).
[57] H. Zhang, S. Si, and C.-J. Hsieh, arXiv preprint arXiv:1706.08359 (2017).
[58] P. Geurts, D. Ernst, and L. Wehenkel, Machine Learning 63, 3 (2006).
SUPPLEMENTARY INFORMATION

S1. SIMULATION METHODS

A. Model

We study a three-dimensional polydisperse mixture of particles interacting via the potential

\[ v(r_{ij}) = \epsilon (\sigma_{ij}/r_{ij})^{12} + \epsilon F(r_{ij}/\sigma_{ij}), \]

only if \( r_{ij} = 1.25\sigma_{ij} \). We use non-additive interactions \( \sigma_{ij} = 0.5(\sigma_i + \sigma_j)(1 - 0.2|\sigma_i - \sigma_j|) \). The function \( F \) is a fourth-order polynomial ensuring continuity of \( v(r_{ij}) \), its first and second derivatives, at the interaction cutoff. The particle diameters \( \sigma_i \) are drawn from the normalized distribution \( P(0.73 < \sigma < 1.62) \propto 1/\sigma^3 \). This model glassformer is efficiently equilibrated with the particle-swap Monte Carlo algorithm, with a fluid robust against fractionation and crystallization down to low temperature [21].

We employ the swap Monte Carlo algorithm implemented in the LAMMPS package, with the optimal parameters provided in [22]. With such parameters supercooled liquid configurations can be generated down to \( T_f = 0.062 \), below the experimental glass transition \( T_g = 0.067 \). In this work, we use configurations prepared in equilibrium conditions at temperatures \( T_f = 0.092, 0.07, 0.062 \).

B. MD dynamics

To explore the energy landscape of glasses generated at \( T_f \), we use classical Molecular Dynamics (MD) simulations. Glasses are first thermalized to \( T_{MD} = 0.04 \) using a Berendsen thermostat. Such temperature is low enough to suppress diffusion, but sufficiently high for quick exploration of the energy landscape. We then run MD simulations in the NVE ensemble, using an integration time step of \( dt = 0.01 \) (LJ time units). Configurations along the MD trajectory are used as the starting point for energy minimization via a conjugate gradient algorithm, which brings them to their inherent structure (IS). MD configurations are minimized every 20, 10, 5 time steps for \( T_f = 0.062, 0.07, 0.092 \) respectively. The values are a compromise between too frequent minimization which sample the same IS many times consecutively, and too infrequent which would miss intermediate IS. For each initial configuration we perform 100, 100, 200 MD runs with different initial velocities for \( T_f = 0.062, 0.07, 0.092 \), respectively. Each run lasts 40000, 100000, 10000 time steps. We performed our calculations on 200, 50, 5 glasses for \( T_f = 0.062, 0.07, 0.092 \). For \( T_f = 0.092 \) glasses we used a subset of the original data obtained in [33].

C. NEB

To analyze the transition between two IS we compute the multi-dimensional minimum energy path separating them, along with its one-dimensional potential energy profile. This is done by the nudged elastic band (NEB) method [49, 50] implemented in the LAMMPS package. We use 40 images to interpolate the minimal energy path, that are connected by springs of constant \( \kappa = 1.0 \epsilon \sigma^{-2} \), and use the FIRE algorithm in the minimization procedure [49, 51].

S2. WHICH MACHINE LEARNING APPROACH SHOULD WE USE?

A. Comparison between ensembled and non-ensembled models

In the main manuscript we discuss how to transform the problem of computing the quantum splitting of two inherent structures into a supervised learning regression problem. Still, many machine learning models can in principle answer this question. We opted for the AutoGluon library [53] which is based on model ensembling to find an optimal solution. Model ensembling consists in averaging the output of different ML models and create an ensemble-averaged model that outperforms each one of its components. In particular, the AutoGluon library performs model ensembling in the form of stacking (stacking different layers of models that are applied in series) and bagging (training multiple instances of the same model on different subsets of data and then combining their prediction).

In Fig. S1(a) we report the most performant ensembling after 20 hours of training, at \( T_f = 0.062, M = 5 \) and \( N_{samples} = 10^4 \). The R2-score reported in the figure shows that the most performant model is a bagged ensemble of CatBoost, a gradient boosting method [54], which then corresponds to the final WeightedEnsemble for this training.
instance. Notice that in the main paper, we always use the best ensembled model. In addition to \textit{CatBoost}, very good performances are also usually achieved by \textit{LightGBM}, another gradient boosting method \cite{57}, and \textit{ExtraTrees}, which is an extremely randomized tree ensemble \cite{58}. Overall, gradient boosting methods typically achieve the highest R2-score. Thus, in the main text when we report results obtained with the ‘best models’, we refer to ensembles of gradient boosting methods assembled with Autogluon.

![Figure S1. Comparison of different ML model ensembles to predict the quantum splitting at $T_f = 0.062$ using $M = 5$ and $N_{\text{samples}} = 7 \cdot 10^4$: (a) R2-score of the test set obtained for an ensemble of models. The full collection was trained for 20 hours of CPU time. In the main manuscript we use the \textit{WeightedEnsemble}. (b) An optimal Neural Network (MLP with 10 hidden layers of size $1.1 \cdot N_{\text{features}}$) does not produce good predictions ($R^2 < 0$).](image)

We motivate our choice of complex ensemble model constructed with Autogluon by showing that simpler models do not perform at the same level. For the system at $T_f = 0.062$ and in the optimal condition corresponding to $M = 5$ and $N_{\text{samples}} = 7000$ the simplest model to at least capture some correlation was a Multi-layer-perceptron (MLP) with 10 hidden layers of size $1.1 \cdot N_{\text{features}}$, using the same input and output structure as in the main text. A first disadvantage of this approach is that several additional hyperparameters have to be selected, such as learning rate, number and size of hidden layers, activation function, etc. We varied these parameters and report in Fig. S1(b) results for the model with the best performance. This optimal MLP was trained until the iteration per epoch was consistently below $10^{-10}$, corresponding to $\sim 10$ hours. Still, the results in Fig. S1(b) demonstrate that the MLP performance is much worse than that of the ensemble models used in the main text. In addition, it is much slower to train. While in this case the predictions could be scaled by a constant to achieve significant improvement, the problem with such an approach is that it is not reliable, and this scaling constant is not known a priori adding an additional layer of complication. For this reason we prefer to use more complex machine-learning models.

### B. Parameters to predict the quantum splitting

A significant advantage of the model ensemble approach is that most of the hyperparameters are automatically optimized by the ensembling. Still, some degrees of freedom have to be fixed. In particular we need to fix the number of particles to be considered ($M$), how many samples to use ($N_{\text{samples}}$) and for how long to train the ensemble. In the next sections we motivate the choices reported in the main text.

- **$M$: input particles** – In Fig. S2(a) we report the effect of $M$, the number of particles considered in the ML procedure. A large $M$ hinders the performance of the ML model because it has to process a larger input vector. The peak performance is achieved at $M = 3$, which is a relatively low number of particles to define a TLS. This confirms that the participation rate in TLS is low \cite{33}.

- **$N_{\text{samples}}$: number of samples** – To evaluate the optimal value of $N_{\text{samples}}$ we report in Fig. S2(b) the variation of the R2-score upon using more samples for training. Notice that we keep the other parameters in the optimal condition for each temperature. The shape of the curves let us conclude that a good number of training samples is $7000, 10000, 30000$ over a total of $14202, 23535, 117370$ for $T_f = 0.062, 0.07, 0.092$ respectively. Notice that
the results depend on the specific quality of the samples provided for training. It is possible to achieve better predictions by selecting better initial samples, more uniformly distributed, which are different from the random choice we make for simplicity. Still, the values reported consistently produce good results, independently from the initial sample composition.

- **Training time** – In Fig. S2(c) we report the effect of the training time on performance. The R2-score approaches the plateau in 10 minutes of training on a single CPU and reaches the final value after $10^3$ minutes ($\sim 16$ hours). Overall, we find that in iterative training, where we retrain the model several times, a good balance between performance and speed can be reached by training for $\sim 10$ minutes. Instead, in a more classic single training approach we suggest to train for $> 10^3$ minutes. Performing the training in parallel can further reduce the training time.

C. Parameters to identify double wells

Figure S3. Accuracy of the double well classifier as a function of (a) $M$, (b) training time, and (c) number of samples. We report the accuracy after training at the lowest temperature $T_f = 0.062$ using the data from Ref. [33].

The ML model has to process a large number of unfiltered minima obtained during the exploration simulation. While we can anticipate that only a small fraction of inherent structure pairs will be a TLS, we also know that the minimum energy path connecting any pair of IS will not form a DW and as it is likely to contain intermediate minima. To exclude those non-DW pairs, we train a classifier using as input the same quantities that we measure to predict the QS, as described in the main text. In Fig. S3 we report the results of a hyperparameter scan at $T_f = 0.062$. These results are similar to those obtained for the QS prediction suggesting that at $T_f = 0.062$ we can use (a) $M = 3$, (b) $10^3$ s of training and (c) $\sim 10^4$ samples in order to achieve $> 95\%$ accuracy, measured over a validation set. We conclude that we can use the same hyperparameters for the DW classifier and the QS predictor.
S3. FEATURES

A. Removing irrelevant features

After finding the optimal parameters to train the ML model, we can extract the specific effect and overall importance of each input parameter, as well as the score drop when each of them is removed. Since each additional feature corresponds to additional computational time and memory, we investigate which features are crucial, and which ones can be excluded to make the ML pipeline faster and more efficient without affecting performance.

![Feature Importance Diagram](image)

**Figure S4. Importance of different features on the quantum splitting predictor.** (a) Performance drop when each specific feature is removed, normalized to the first most important. Different colors correspond to different glass preparation temperatures. (b) Shapley values calculated at $T_f = 0.062$. In general, at any temperature the most important feature is the energy difference ($\Delta E$) of the two IS, which is followed by the value of the displacement of the $M$ particles. The Shapley values on the right also report the effect of the specific features: the splitting $\Delta E$ has to be small (green) in order to predict low QS (i.e. negative SHAP), while instead the displacements have to be large (red) in order to point towards low QS.

In Fig. S4(a) we report the feature importance (in log scale) after a single iteration of training containing all the data from Ref. [33]. Different colors refer to glasses prepared at different temperatures. Independently from the temperature, the most important features are the energy difference $\Delta E$ between the two IS, followed by the total displacement of particles $\sum_i \Delta r_i$. The third most important information is the positions of the particles that displaced the most $\sum_i |\Delta r_i|$. After them, we find that the arrangement of the first shell of neighbors represented by the $q$ parameters and the particle sizes $\sigma_i$ are insignificant (and so is their variation between the two IS of the pair $\Delta$ reported in Fig. S4), since their importance is more than two orders of magnitude smaller than the one of the energy and displacements. Since the $q$ parameters are also the slowest to compute, we decided to not include them in the final pipeline reported in the main manuscript.

Next, to quantify the role of those inputs, we calculate their Shapley values [56] shown in Fig. S4(b) for $T_f = 0.062$. The color codes for the value taken by the specific feature (red when the value is high, green when low) and reports its scaled impact on the model output. The x axis reports the Shapley value, where $\text{SHAP} > 0$ implies that the specific feature is pushing towards predicting a high QS, while $\text{SHAP} < 0$ indicates that the feature promotes low QS values. While no input feature alone can predict a TLS (because $|\text{SHAP}|$ is small) there are some visible trends: (i) the energy asymmetry is the most important feature and should be small in order to predict a low QS, (ii) the particle displacements have to be larger than a threshold in order to predict low QS.

According to these results we rationalize the ML prediction. The energy difference between two IS is the main predictor for the quantum splitting, which is never too large for DW, then the displacements are necessary to understand if the two IS are similar and what their stability is (see the temperature classification section in the SI). The displacement nucleus size complements the information contained in the displacements and gives local information about the participation ratio. Finally, all the other features do not provide any improvement. Since the bond order parameters are computationally expensive to calculate, we do not include them in the final ML approach we propose in the main text. The performance reported in the main text confirms that our choice is justified.
B. Reducing the number of features

We justify the choice of features discussed in sec. IVa and Fig. 2 by presenting the performance of our ML model as a function of feature number. In Tab. S1 we rank the features according to their Shapley values (see Fig. 6). We report (blue line) in Fig. S5 the accuracy of the DW classifier (a) and the QS predictor (b) as a function of number of features, following the Shapley ranking of Tab. S1. The DW classifier reaches its best accuracy with six features. In the initial study Ref. [33, 36], the matrix $T$ was the only information used to analyze the pair of IS. Here, the classifier already reaches $\sim 90\%$ accuracy using the transition matrix only (two features). We add (orange line) the performances when we exclude $\Delta E, T_{ij}$ and $T_{ji}$, which are the most important features overall. Surprisingly, the DW classifier still reaches its maximum accuracy with two features ($\Delta r_2$ and $|r_0 - r_2|$) as highlighted in the inset. The QS predictor shows instead very good predictions even using a single feature ($\Delta E$ in the blue curve, or $\Delta r_0$ in the orange).

| Shapley ranking | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| DW classification | $T_{ij}$ | $T_{ji}$ | $\Delta r_2$ | $|r_0 - r_2|$ | $|r_0 - r_1|$ | $d$ | $\Delta r_0$ | $\Delta E$ | PR | $\Delta r_1$ |
| QS prediction | $\Delta E$ | $\Delta r_0$ | $d$ | $\Delta r_1$ | $|r_0 - r_1|$ | $|r_0 - r_2|$ | $\Delta r_2$ | $T_{ij}$ | $T_{ji}$ | PR |

Table S1. Ranking the features used in the main manuscript by their Shapley values.

Figure S5. Effect of the number of features. Accuracy of the DW classifier (a) and Pearson correlation score of the QS predictor (b), as a function of the number of features used. Along the blue curves the features are ranked by their importance using their Shapley values, so the first features to be used are the most important. For the orange curve we exclude $\Delta E, T_{ij}$ and $T_{ji}$ to evaluate the performances of the ML model without the features with the best Shapley values.

S4. ITERATIVE TRAINING

The standard supervised learning approach consists in collecting a significant amount of data, then using them to train the ML model and finally use the model predictions. While overall very powerful, this scheme is not optimal when the goal of the model is to drive the exploration in a space much larger than the available data. This is the case of our main study where we employ the ML model to identify IS pairs with low QS. In order to make our approach more efficient and generalizable we developed the iterative training procedure.

We start the iterative training procedure from a sample of $K_0$ pairs for which we calculate the QS. We empirically find that $K_0 \sim 5000$ achieves a good balance between precision and time. This sample has to be balanced between DW and non-DW, but there is no need to include TLS in the initial sample. It is possible to start from a smaller $K_0$, at the cost of a performance drop in the first few iterations of training. From the initial sample we perform a first training that takes 10 minutes for the DW classifier and 10 minutes for the QS predictor. We then use the model to predict the $K_i = 500$ IS pairs with the lowest QS. We report the cumulative number of TLS, DW, and non-DW at each iteration.
Figure S6. Performance of the iterative training procedure at $T_f = 0.062$ for the data collected in the main manuscript. Starting from $K_0 = 5000$ samples we perform iterations of $K_i = 500$ predictions for 11 iterations. In (a) we report the cumulative number of double wells (DW), two-level systems (TLS) and non double-wells (Non-DW). In (b) we break down this result by color coding the confusion matrix of the ML model at each iteration of iterative training.

In Fig. S6(a). From Fig. S6(b) we see that during the first iteration most of the $K_i = 500$ best candidates are actually non-DW, so the model is performing poorly. This is the reason why we suggest to perform only $K_i = 500$ NEBs for each iterations, otherwise the first iteration would lead to wasting time to perform calculations over non-DW. On the other hand, in the first iteration 72 TLS are already found by running 500 NEBs. This is to be compared with Ref. [33] in which 61 TLS were found by running $>14000$ NEBs. After the first retraining (during iteration 2) the performance of the ML model is already excellent and less than 30% of the 500 best pairs are non-DW, while 134 are newly found TLS.

We also used our iterative training to reprocess the data of Ref. [33], and we report the results in Table 1 (main text). We show in Fig. S7 a detailed analysis of the procedure at $T_f = 0.062$. While the standard approach was able to identify 61 TLS running $>14000$ NEB+Schrödinger calculations, iterative training finds 156 TLS, by running only 2500 NEBs. This confirms that more than half of the total TLS were hidden among the pairs discarded by Ref. [33]. In details, Fig. S7(a) shows the cumulative number of DW and TLS that we find, while Fig. S7(b) reports the confusion matrix for the different steps of iterative training.

Figure S7. Reprocessing the data of Ref. [33] at $T_f = 0.062$ with iterative training. In (a) we report the cumulative number of double wells (DW), two-level systems (TLS) and non double-wells (Non-DW). While Ref. [33] (red dashed line) identified 61 TLS running $>14000$ NEBs, iterative training finds 156 TLS from 2500 NEBs. In (b) we report the confusion matrix of the 5 steps of iterative training that we run.
Overall, these results demonstrate that the ML approach is not only faster than a manual filtering rule based on the transition matrix, but also much more effective. The pool of IS pairs excluded from the analysis in the original approach effectively contains a significant number of TLS. In conclusion, a ML driven exploration is not only more efficient than manual filtering, but also necessary to capture the correct statistics.

S5. TEMPERATURE AND STABILITY

Our ML approach is able to predict the quantum splitting of a pair of IS from static information. It is known that TLS have different features depending on their preparation temperature, or glass stability [33]. Here we address the following questions: (i) how difficult is it for a machine to distinguish data corresponding to different temperatures, (ii) what are the most important features for this task, and (iii) does our ML approach learn temperature-independent features?

A. Temperature classification

To understand how TLS and IS features evolve with temperature, we trained a multi-layer-perceptron (MLP) to classify the temperature corresponding to a specific pair. We find that it is possible to rapidly train a classifier reaching accuracy > 95%. Depending on the value of $M$, the input layer is composed by $N_{\text{features}}$ neurons, where the features are explained in the main text. After the input layer, there are $n$ hidden fully connected layers, all of the same size $s$. The activation function for each neuron-neuron connection is a ReLu function. The last layer is composed of 3 neurons that represent the probability that a given input belongs to one of the three classes: $T_f = 0.062$ or $T_f = 0.07$ or $T_f = 0.092$. The performance of the MLP are evaluated by measuring the accuracy, which is the percentage of the corrected predictions. In Fig. S8(a) we report the effect of increasing the size of the hidden layers, while in Fig. S8(b) we report the effect of a larger number of hidden layers.

![Figure S8](image)

Figure S8. Temperature classifier accuracy as a function of number of hidden layers (a) and their size (b). We report results for $M = 5$. Results suggest that already with 2 hidden layers of 60 neurons it is possible to achieve an accuracy above 95%.

Our results show that a MLP with 2 hidden layers of 60 neurons has a 95% accuracy after less than 1h of supervised training. The answer to question (i), is that one can identify the temperature at which a pair of IS was obtained.

B. Static signatures of glass stability

We answer question (ii) by measuring the Shapley (SHAP) values from the temperature classifier. In Fig. S9 (a) we report the SHAP value of the most important input features for the MLP prediction that considers only the $M = 5$ particles that displaced the most. The smallest ($i = (M - 1) = 4$) and the largest ($i = 0$) particle displacements emerge as the most important input, noted $\Delta \rho_4$ and $\Delta \rho_0$, respectively. Their average effect on the model prediction...
is shown in Fig. S9(b,c). The particle that displaced the most ($i = 0$) shows a large displacement at low preparation temperature, while the particle that displaced the least ($i = M - 1$) exhibits a large displacement at high temperature. Our results suggest that higher temperatures are characterized by more collective rearrangements, leading to large $\Delta r_4$. In glasses prepared at low temperature instead, transitions are characterized by a particle displacing significantly more than the others.

Figure S9. Microscopic differences originating from different temperature preparation/glass stability, quantified using the Shapley values for the temperature classifier. (a) Summary of the SHAP values for the 9 most important input features, measured for a subset of 1000 pairs processed by the $T$-classifier. Predicted glass preparation temperature as a function of (b) $\Delta r_4$, and (c) $\Delta r_0$, with all other inputs taking their average value.

C. Transferability and crossvalidation

We address question (iii) by testing how our model performs when trained at $T_{\text{train}}$ and deployed to predict the quantum splitting of pairs at $T_{\text{predict}} \neq T_{\text{train}}$. In Fig. S10 each column corresponds to a training temperature: $T_{\text{train}} = 0.092$ (left), 0.07 (middle) and 0.062 (right). Then, computed quantum splittings are compared with ML prediction made on samples obtained at $T_{\text{predict}}$, decreasing from top to bottom. We see in Fig. S10 that the predictive power of a model trained at a different temperature is slightly lower compared to the model trained at the same temperature (Fig. 3 of main). Still, the transferability of the model is good, especially when the model is trained at $T_{\text{train}} > T_{\text{predict}}$. This situation is the most useful, since it is easier to obtain data at higher $T$. We conclude that if not enough data is available at low temperature, it is possible to rely on model transferability. This relies on the fact that TLS share some general temperature-independent features.

In summary, we have seen that IS and TLS have different microscopic features when generated from different stabilities. It is then optimal to train the ML model at a fixed temperature only. If no easier way to measure $T$/stability are available, it is possible to use ML to classify the stability of each IS by adding another block to the workflow, as reported in the main manuscript. Alternatively, at the cost of a finite accuracy drop it is even possible to transfer the model predictions at different temperatures and thus train where data collection is fast and easy and apply it where it is not.
Figure S10. Transferability and crossvalidation of the machine learning model. Exact quantum splitting against ML prediction. The quantum splitting predictor is trained at \( T_{\text{train}} \) and deployed at \( T_{\text{prediction}} \neq T_{\text{train}} \). From left to right: \( T_{\text{train}} = 0.092, 0.07, 0.062 \). From top to bottom, decreasing \( T_{\text{prediction}} \). Performances are not on par with Fig. 3, but good in particular when \( T_{\text{train}} > T_{\text{predict}} \), which is of practical interest.