nature portfolio

https://doi.org/10.1038/s41567-021-01201-7

Supplementary information

Learning models of quantum systems from experiments

In the format provided by the authors and unedited

Supplementary Information
Learning models of quantum systems from experiments

Antonio A. Gentile, Brian Flynn, Sebastian Knauer, Nathan Wiebe, Stefano Paesani,
Christopher E. Granade, John G. Rarity, Raffaele Santagati, Anthony Laing

1. Additional details about the QMLA protocol

A. Concepts

We review here all the concepts introduced with the description of Quantum Model Learning Agent (QMLA) in the main text.

- ***Models***: individual candidate Hamiltonian models ($\hat{H}_j$). In particular for the instances studied in this paper. Note how we often adopt a compact representation, e.g. $\hat{S}_{xy} = \alpha_x\hat{\sigma}_x + \alpha_y\hat{\sigma}_y$, $\hat{S}_y\hat{A}_x = \alpha_y\hat{\sigma}_y + \alpha_{xx}\hat{\sigma}_x \otimes \hat{\sigma}_x$.
- ***Directed Acyclic Graphs*** retain all the information gathered by the QMLA. Nodes represent models after undergoing QHL ($\hat{H}'_j$). Edges in the structural directed graph (sDG) represent the relationships (parent/child) between those models. Edges in the comparative directed graph (cDG) between nodes (i, j) give the pairwise BFs, $\mathcal{B}_{i,j}$.
- ***Layers***: a group of nodes μ, belonging to the same generation, and hence forming a clique in the cDG. Layers usually consist of similar models, e.g. they span the Hilbert spaces of the same dimension, and differ only by a single term. For instance, in Fig. 2f (main text) the fourth layer $\mu^4 = \{\hat{S}_{xyz}\hat{A}_x, \hat{S}_{xyz}\hat{A}_y, \hat{S}_{xyz}\hat{A}_z\}$.
- ***Consolidation***: Comparison between any set of models $\mathbb{H}_k = \{\hat{H}_k\}$, achieved by computing a pairwise metric of choice $\forall \hat{H}_i, \hat{H}_j \in \mathbb{H}_k$. E.g. in Fig. 2f we use $\mathcal{B}_{i,j}$. Each pairwise comparison for $\hat{H}_j$ gains that model a point, summing to the score s_j, i.e. the *indegree* of the node $\hat{H}_j$. Consolidation provides a ranking of models in $\mathbb{H}_k$. When consolidation is performed upon a layer μ, the highest ranking model(s) are named the ***layer champion***(s), $\hat{H}_{C(\mu)}(\{\hat{H}_{C(\mu)}\})$. When instead consolidation spans the whole directed graph (DG), the highest ranking model is the nominated global champion $\hat{H}'$.
- ***Pruning***: deactivation of a model, such that the model is not entertained anymore for following consolidation or spawning. It can be thought of as an effective pruning from the graph. For all QMLA instances reported in this paper, following consolidation within layer μ, all non-champion models are always pruned.
- ***Primitives***: a set of base models used to generate candidate models. Primitives should be a physically interpretable, basic set of operators, and are user-defined. An obvious possible choice is the Pauli set $\{\hat{\mathbb{1}}, \hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z\}$. Summation and tensor product operations among them grants access to potentially any Hamiltonian of interest for the characterisation of quantum devices [1].
- ***Spawning***: QMLA explores the space of models by trying various combinations of the primitive terms. In particular, this is done by spawning new models from ***seed*** models (see the groth rules below). For instance, found a layer champion $\hat{H}_{C(\mu)}$, models belonging to a new layer $\nu(= \mu + 1)$ may be spawned using $\hat{H}_{C(\mu)}$as parent node.
- ***Exploration strategies*** describe how many models per layer are retained as seed, and how to combine primitives to generate a new set of models. These rules are also user-defined, but different rules can inform QMLA with a different amount of a priori structure.
  - As already outlined in Methods, here we adopt as exploration strategy (ES) either a *greedy search* or a *genetic algorithm*.

  - the ES includes a definition of a *termination rule*, i.e. a test flagging when QMLA should not spawn any further layer, but consolidate across the existing ones.

- ***Layer collapse***: upon the termination rule, a preliminary inter-layer comparison is conducted, whereby $B_{C(\mu),C(\nu)}$ are computed solely for neighbouring layers, i.e. between parent $\hat{H}_{C(\mu)}$and child $\hat{H}_\nu$ nodes. In case of strong evidence favouring the parent (child) layer, i.e. if $B_{C(\mu),C(\nu)} \gg 1$ ($\ll 1$), the losing model - and hence its entire layer - is pruned. The inheritance rule, in this case, dictates that the grandchildren (children) node(s) will inherit terms from its grandparent instead.

B. Protocol Steps

In the same way, we review here the fundamental steps of the QMLA protocol in full generality. Further details for the specific implementation(s) adopted in this paper are reported in the pseudocodes, § 1 D.

1. Initialise the sDG. When there is no a priori information available for a bootstrap, the first layer includes the primitives alone.
2. Iteratively, and until the termination rule occurs (denoted the most recently added layer as μ):
   (a) Train all models in μ by performing a parameter estimation method of choice, $\forall \hat{H}_j \in \mu$
   (b) Consolidate μ and select the layer champion(s).
   (c) Spawn a new layer: by using the layer champion(s) as seed, use the ESs and primitives to generate a new set of children models in layer ν.
3. Perform comparisons only between parent/children pairs that are both layer champions.
4. Eventually enact layer collapses, obtaining a reduced set $\mathbb{H}'_C$.
5. Consolidate $\mathbb{H}'_C$ to find the global champion model, $\hat{H}'$.
6. Attempt a reduction of $\hat{H}'$.

We briefly comment about the last point alone, as the remaining has already been discussed before. Once $\hat{H}'$is nominated, it is possible to adopt an additional step to counter overfitting phenomena (beyond what is already achieved by adopting BFs), which might arise e.g. because of the greediness in the ES, see § 1 F. This *reduction* step is as follows: if the probability distribution for some parameters is found to be within one standard deviation from 0, QMLA proposes a reduced model, $\hat{H}'_r$, which removes the corresponding terms. We then compute the Bayes factor, $\mathcal{B}(\hat{H}'_r, \hat{H}')$: if $\mathcal{B} > 100$, then we prefer the simpler model. We emphasise how the threshold $\mathcal{B} > 100$ is stringent enough, such that any non-negligible contribution to the predictive power will not be discarded.

C. Quantum Hamiltonian Learning

The interested reader can refer for details about the Quantum Hamiltonian Learning (QHL) protocol to [2–4], and for its experimental demonstration to [5, 6]. Here we thus focus solely upon reporting details of this QHL implementation, that slightly differ from the original proposal.

1. *Details about the parameter estimation technique*

In order to learn Hamiltonian parameters we employ QHL, a protocol assuming access to a trusted (quantum) simulator to (efficiently) test trial Hamiltonians forms $\hat{H}_j$ and their corresponding parametrisations $\mathbf{x}_j$ [2, 3, 5]. The Bayesian inference in QHL is here performed via an approximate technique known as sequential Monte Carlo (SMC), which encodes the knowledge about the parameters in a probability distribution P, discretised in a set of *particles* $\{\mathbf{x}_i\}$. The implementation of SMC and of a Liu-West resampler [7] is here obtained leveraging upon the `Qinfer` python package [8]. In particular, we keep the resampling parameters $a = 0.98, l_{res} = 0.5$. Resampling redraws particles according to the updated probability distribution, focusing the limited available particles in those regions where the support of P is higher.

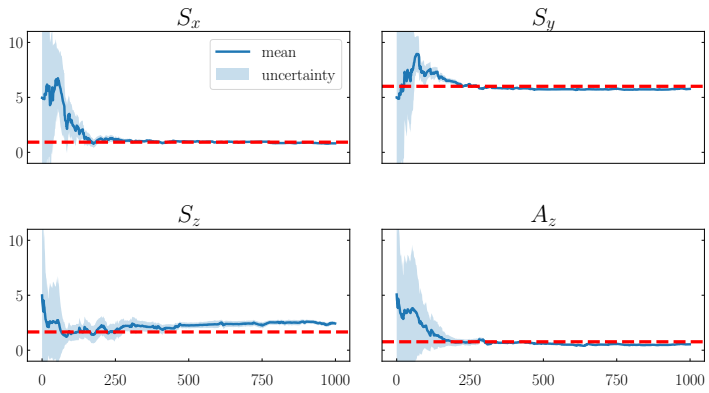


Figure S1 Parameter learning for $\hat{S}_{xyz}\hat{A}_z$ using the parameters used during QMLA in the main text. Each parameter is learned simultaneously for $N_E = 1000, N_P = 5000$; the probability distribution (projected onto single parameters' space) is shown in light blue, with the mean of the distribution in dark blue. In each case QLE directs the particle distribution towards the true parameters (red).

A key step in QHL is extracting data from the systems and comparing those with the likelihoods computed by the simulator interfaced with the system itself. In both cases, a probe initial state $|\psi\rangle$ is evolved for a time t, before operating a projective measurement, whose outcome is here the binary datum d (however, the simulator can of course compute the full likelihood function $\mathcal{L}(d|\mathbf{x}_i, t, |\psi\rangle, \hat{H}_j)$ for each particle $\mathbf{x}_i$). That is, likelihoods range in $[0, 1]$ and reflect the prediction of observing datum d, for a chosen $\mathbf{x}_i$. For a successful learning instance across N_e epochs (each characterised by a single *experiment* on the system and an update step for $P(\mathbf{x})$), we expect P to progressively narrow towards the true parameter vector. In a typical multidimensional parameter space, this can be observed upon convergence via the marginals of the posterior distribution, that we plot for an exemplary case in Fig. S1. A useful concept describing this is the *volume* V of the posterior distribution: we can envision a particle cloud in high dimensional space encoding $P(\mathbf{x})$, which can be enclosed e.g. by an ellipsoid. When the posterior distribution shrinks towards the (correct) value, the volume of said ellipsoid shrinks as well, characterising quantitatively the particles' convergence. V is also renormalised to take into account the size of different parameterisations $|\mathbf{x}|$, and hence their $|\mathbf{x}|$-dimensional space of reference. V was indeed used as a metric in Fig. 2d. From the same figure it is evident how models of different complexity (in terms of size of their parametrisation) will learn at different rates, as more parameters need to converge towards the correct value. In order to allow enough time for the most complex models, without spending too many epochs with the simplest ones, we envisage how an efficient way to halt the QHL subroutine for each $\hat{H}_j$ is to assign N_e dynamically. That is, we can check periodically for saturation in the reduction of V_j, as a signature of a converged QHL instance.

2. *Experiment design and heuristics*

The backbone of QMLA is thus to perform QHL, for each explored model $\hat{H}_j$, for a variable number of epochs N_e, obtaining cumulatively a set of observed data D_j from the quantum system, and updating the parameters $\mathbf{x}_j$ accordingly. When launching QHL instances, we assume that the QMLA is able to access key experimental controls in the system to characterise: the evolution time t and the initially prepared probe state for the system $|\psi\rangle_{\text{sys}}$ (see Fig. 1e). However, typically in QHL the system is assumed closed, and hence its evolution entirely described by a $\hat{H}_{\text{sys},0}$, and its current state by $|\psi\rangle_{\text{sys}}$.

This need not be always the case when characterising an unknown quantum system, that might be open and therefore affected by the contribution of an external environment. By instance, this is the case of all the experiments with NV centres reported here. In such cases, the effective Hamiltonian governing the observed evolution can be written as [9]: $\hat{H}_{\text{glo},0} = \hat{H}_{\text{sys},0} + \hat{H}_{\text{env},0} + \hat{H}_{\text{int},0}$, with $\hat{H}_{\text{int},0}$ describing the interaction between system and environment. In order to address this more general task, we introduce the following approach. First, for all the results in this paper we prepare states $|\psi\rangle_{\text{sys}}$ in the quantum system independently from its

environment, so that input states are separable, of the form:

$$|\psi\rangle = |\psi\rangle_{\text{sys}} \otimes |\phi\rangle_{\text{env}} \,. \tag{S1}$$

where for brevity, we omit the subscript for global states. The different $|\psi\rangle$ adopted for simulations and experiments have been listed in Methods. Eq. S1 takes into account how, in most cases, the environment will not be directly accessible. For the same reason, likelihood functions obtained as:

$$\left| \left\langle d \middle| \exp\left(-i\hat{H}_{\text{sys},j}(\mathbf{x}_j)\ t\right) \middle| \psi \right\rangle_{\text{sys}} \right|^2 \tag{S2}$$

are in general meaningless, and a preliminary partial trace over the environmental degrees of freedom needs to be taken into account. The likelihood for the Bayesian inference is thus here modified as:

$$\mathcal{L}(d|\hat{H}_{\text{glo},j}, \mathbf{x}_j, t, |\psi\rangle) = \langle d|\text{tr}_{\text{env}}\left(\rho_{\text{glo}}\right)|d\rangle_{\text{sys}} \,, \tag{S3}$$

with $|d\rangle_{\text{sys}}$ the chosen measurement basis for $\mathcal{H}_{\text{sys}}$, and

$$\rho_{\text{glo}}(\hat{H}_{\text{glo},j}, \mathbf{x}_j, t, |\psi\rangle) = \exp\left(-i\hat{H}_{\text{glo},j}t\right) |\psi\rangle\langle\psi| \exp\left(i\hat{H}_{\text{glo},j}t\right) \tag{S4}$$

the global density matrix evolved under each $\hat{H}_{\text{glo},j}$.

The remaining control, evolution time t, must be designed in order to maximise the information gained from the experiments: at short t, the discrepancies between coarsely trained $\mathbf{x}_j$ are intuitively expected to be less evident, or might become negligible coincidentally. This is exemplified e.g. by Fig. S3, where even wrong models can return an approximately faithful dynamics for $t < 1\mu s$. Therefore, it is clear how to fully characterise the system, the model ought to train on higher t also, where also decoherence phenomena become evident. We adopted here a heuristic method followed by a schedule: i) first, we dynamically choose t at a pace dictated by the information gained to date; ii) then we force the training to use a linear schedule to exhaust all the points in the dataset. The second step is fundamental to avoid so that it must attempt to address complex dynamics. Nevertheless, such fixed schedules might well be non-optimal in general.

More in detail, the first phase adopts the *particle guess heuristic* (PGH) [2, 10]), i.e. t is chosen as the reciprocal of the $L1$ distance $\sum_{i=1}^{n} \left|\mathbf{x}^1 - \mathbf{x}^2\right|$ between two particles $\mathbf{x}^1, \mathbf{x}^2$ sampled from the posterior. Intuitively, if the volume is large, the distance between two randomly sampled particles will be large, so t will be small, ensuring QHL focuses on small t when the confidence in the learned parameters is low, and viceversa.

D. Pseudocode for QMLA

We provide pseudocode for the main QMLA routine, Alg. 1, as well as key subroutines for QHL that have been modified here ad-hoc in Alg. 2, and the calculation of Bayes factors, Alg. 4. Finally, subroutines for consolidating models are in Alg. 3-5. Some functionality is assumed within the pseudocode, in particular operations on the DG to extract information such as parentage, or to prune models, though these cases are self explanatory from the given function names.

Algorithm 1: Quantum Model Learning Agent

Input: DG // Directed acyclic graph, with structural and comparitive components
Input: Train // callable model training subroutine, e.g. Alg. 2
Input: Consolidate() // callable function, Alg. 3
Input: $Q, N_E, N_P, f(P(\mathbf{x}))$ // as in QHL
Input: $P(\mathbf{x})\ \forall \mathbf{x} \in \mathbf{X}$ // probability distribution for any parameterisation $\mathbf{x}$ within the space of valid Hamiltonian parameterisations $\mathbf{X}$.
Input: $\mathbb{H}_0$ // set of primitive models from which to construct Hamiltonian models
Input: R // rule(s) by which to generate new models
Input: f_T // function to check whether to terminate QMLA
Output: $\hat{H}^C$ // Approximate Hamiltonian model form

$\mathbb{H}^c \leftarrow \{\}$ // Initialise champion model set
$\mu \leftarrow 0$ // First layer
$\mathbb{H}^\mu \leftarrow R(\mathbb{H}_0)$ // Get first batch of models from exploration strategy and primitives

while f_T*(DG) == False* **do**
  if μ *!= 0* **then**
    $\mathbb{H}^\mu \leftarrow R(H_C^{\mu-1})$ // Generate new set of models using previous layer champion as seed
  end
  for $\hat{H}_k \in \mathbb{H}^\mu$ **do**
    Train$(Q, \hat{H}(\mathbf{x}_k), P(\mathbf{x}_k), N_E, N_P) \leftarrow \mathbf{x}'_k$ // optimal parameter set for model $\hat{H}_k$
    $\hat{H}_k \leftarrow \hat{H}(\mathbf{x}'_k)$
    DG.update$(\hat{H}_k)$ // Update learned model on DG within QMLA class
  end
  $\hat{H}_C^\mu \leftarrow$ Consolidate$(\mathbb{H}^\mu)$
  $\mathbb{H}^C \leftarrow \{\mathbb{H}^C, \hat{H}_C^\mu\}$ // Add layer champion to champion set
  $\mu \leftarrow \mu + 1$
end
$\mathbb{H}^C \leftarrow$ parentalConsolidation(DG)
$\hat{H}^C \leftarrow$ Consolidate$(\hat{\mathbb{H}}^{\mathbb{C}})$ // Assign global champion model as champion of champion model set

return $\hat{H}^C$ // Return champion model

Algorithm 2: Quantum Hamiltonian Learning (exemplary `Train` subroutine)

Input: Q `// some physically measurable or simulatable quantum system to study.`
Input: $\hat{H}'$ `// Hamiltonian model with which to attempt reconstruction of data from` $\hat{H}_0$
Input: $P(\mathbf{x})$ `// probability distribution for` $\mathbf{x} = \mathbf{x}_0$
Input: N_E `// number of epochs to iterate learning procedure for`
Input: N_P `// number of particles (samples) to draw from distribution at each epoch`
Input: `RS()` `// Resampling algorithm with which to redraw probability distribution`
Output: $\mathbf{x}'$ `// estimate of Hamiltonian parameters`

Sample N_P times from $P(\mathbf{x}) \leftarrow$ particles
for $e \in \{1 \rightarrow N_E\}$ **do**
  Sample x_1, x_2 from $P(\mathbf{x})$
  $t \leftarrow \frac{1}{||x_1 - x_2||}$
  for $p \in \{1 \rightarrow N_P\}$ **do**
    Retrieve particle $p \leftarrow \mathbf{x}_p$
    Measure Q at $t \leftarrow d$ `// datum`
    $|\langle\psi| e^{-i\hat{H}(\mathbf{x}_p)t} |\psi\rangle|^2 \leftarrow E_p$ `// expectation value`
    $Pr(d|\mathbf{x}_p; t) \leftarrow l_p$ `// likelihood, computed using` E_p `and` d
    $w_p \leftarrow w_p \times l_p$ `// weight update`
  end
  if $1/\sum_p w_p^2 < N_P/2$ `// check whether to resample (are weights too small?)`
  then
    $P(\mathbf{x}) \leftarrow$ `RS`$(P(\mathbf{x}))$ `// Resample according to provided resampling algorithm`
    Sample N_P times from $P(\mathbf{x}) \leftarrow$ particles
  end
end
`mean`$(Pr(\mathbf{x})) \leftarrow \mathbf{x}'$
return $\mathbf{x}'$

Algorithm 3: Layer Consolidation

Input: $\mathbb{H}$ `// set of models to consolidate`
Input: N `// number of top-ranked models to return; default 1`
Input: b `// threshold for sufficient evidence that one model is stronger`
Input: `BF()` `// callable function to compute the Bayes factor between` $\hat{H}_j$ `and` $\hat{H}_k$`, Alg. 4`
Output: $\hat{H}_C$ `// Superior model among input set`
for $\hat{H}_j \in \mathbb{H}$ **do**
  $s_j = 0$ `// Initialise score for each model`
end
for $\hat{H}_j, \hat{H}_k \in \mathbb{H}$ `// pairwise Bayes factor between all models in the set`
do
  $B \leftarrow$ `BF`$(\hat{H}_j, \hat{H}_k)$
  if $B \gg b$ `// Increase score of winning model`
  then
    $s_j \leftarrow s_j + 1$
  else if $B \ll 1/b$ **then**
    $s_k \leftarrow s_k + 1$
end
$k' \leftarrow \max_k\{s_k\}$ `// Find which model has most points`
return $\hat{H}_{k'}$

Algorithm 4: Bayes Factor calculation

Input: Q // some physically measurable or simulateable quantum system to study.
Input: $\hat{H}'_j, \hat{H}'_k$ // Hamiltonian models after QHL (i.e. $\mathbf{x}_j, \mathbf{x}_k$ already trained), on which to compare performance.
Input: T_j, T_k // set of times on which $\hat{H}'_j$ and $\hat{H}'_k$ were trained during QHL, respectively.
Output: B_{jk} // Bayes factor between two candidate Hamiltonians; describes the relative performance at explaining the studied dynamics.

$T = \{T_j \cup T_k\}$
for $\hat{H}'_i \in \{\hat{H}'_j, \hat{H}'_k\ \}$ **do**
  $l(D|\hat{H}_i) = 0$ // total log-likelihood of observing data D given $\hat{H}_i$
  for $t \in T$ **do**
    Measure Q at $t \leftarrow d$ // datum for time t
    Compute $|\langle\psi| e^{-i\hat{H}'_i t} |\psi\rangle|^2 \leftarrow E_p$ // expectation value for $\hat{H}'_i$ at t
    Compute $Pr(d|\hat{H}_i, t)$ // from Bayes' rule, using d, E_p
    $log\left(Pr(d|\hat{H}_i, t)\right) \leftarrow l$ // likelihood of observing datum d for $\hat{H}'_i$ at t
    $l(D|\hat{H}'_i) + l \leftarrow l$ // add this likelihood to total log-likelihood
  end
end
$exp\left(l(D|\hat{H}'_j) - l(D|\hat{H}'_k)\right) \leftarrow B_{jk}$ // Bayes factor between models

return B_{jk}

Algorithm 5: parentalConsolidation

Input: DG // Directed acyclic graph with information about nodes' relationships/comparisons
Input: BF() // callable function, Alg. 4
Input: b // threshold for sufficient evidence that one model is stronger
Output: $\mathbb{H}^C$ // Superior model among input set

DG.surviving_layer_champions() $\leftarrow \mathbb{H}$
for $\hat{H}_c \in \mathbb{H}$ **do**
  DG.parent($\hat{H}_c$) $\leftarrow \hat{H}_p$
  $B_{pc} \leftarrow BF(\hat{H}_p, \hat{H}_c)$
  if $B_{pc} > b$ **then**
    DG.prune($\hat{H}_c$)
  else if $B_{pc} < 1/b$ **then**
    DG.prune($\hat{H}_p$)
end
DG.surviving_layer_champions() $\leftarrow \mathbb{H}$
return $\mathbb{H}$

Algorithm 6: Exploration strategy: genetic algorithm

Input: ν // information about models considered to date

Input: $g(\hat{H}_i)$ // objective function which assess models $\hat{H}_i$

Output: $\mathbb{H}$ // set of models

$N_m = |\nu|$ // number of models
for $\hat{H}_i \in \nu$ **do**
  $f_i \leftarrow g(\hat{H}_i)$ // model fitness via objective function
end
$r \leftarrow \texttt{rank}(\{f_i\})$ // rank models by their fitness
$\mathbb{H}_t \leftarrow \texttt{truncate}(r, \frac{N_m}{2})$ // truncate models by rank: only keep $\frac{N_m}{2}$
$s \leftarrow \texttt{normalise}(\{f_i\})\ \forall \hat{H}_i \in \mathbb{H}_t$ // normalise remaining models' fitness
$\mathbb{H} = \{\}$ // new batch of chromosomes/models
while $|\mathbb{H}| < N_m$ **do**
  $p_1, p_2 = \texttt{roulette}(s)$ // use s to select two parents via roulette selection
  $c_1, c_2 = \texttt{crossover}(p_1, p_2)$ // produce offspring models
  $c_1, c_2 = \texttt{mutate}(c_1, c_2)$ // probabilistically mutate
  $\mathbb{H} \leftarrow \{\mathbb{H}, c_1, c_2\}$ // add new models to batch
end
return $\mathbb{H}$

E. Stability analysis for QMLA

The principal question addressed in this section is the stability of Bayes factor analysis under the sequential Monte-Carlo approximation. BFs are known to be a statistically robust measure for comparing the predictive power of different models, whilst favourably weighting less structure to limit overfitting [11], and have been successfully used in the context of resolving multi–modal distributions in [12]. While stability has broadly been shown in the past for the problem of quantum Hamiltonian learning [2–4], the problem of using BFs to compare alternative families of models is not directly covered by those results. The ultimate reason to adopt a logarithmic likelihood within BFs (Eq. 1), was that likelihoods are prone to numerical instabilities as the number of sample measurement grows [8], and hence adopting a log-difference instead of a ratio can help reducing artefacts. Here we claim that adopting BFs to our purposes is legitimate, provided that both the prior distribution and the likelihood function are sufficiently smooth functions of the Hamiltonian coefficients.

Let $\epsilon \leq x/2$ and $\delta \leq y/2$ then

$$\begin{aligned}\left|\frac{x+\epsilon}{y+\delta} - \frac{x}{y}\right| &\leq \left|\frac{x+\epsilon}{y+\delta} - \frac{x+\epsilon}{y}\right| + \left|\frac{x+\epsilon}{y} - \frac{x}{y}\right| \\ &\leq \frac{|\epsilon|}{|y|} + |x+\epsilon|\left|\frac{1}{y} - \frac{1}{y+\delta}\right|. \\ &= \frac{|\epsilon|}{|y|} + \frac{|x+\epsilon|}{|y+\delta|}\frac{|\delta|}{|y|} \\ &\leq \frac{|\epsilon|}{|y|} + 3\frac{|x|}{|y|}\frac{|\delta|}{|y|}. \end{aligned} \tag{S5}$$

The Bayes factor for two models is defined to be of the form, given data D is observed under experimental settings E, is

$$\mathcal{B} := \frac{P(D|M_1;E)}{P(D|M_2;E)} \tag{S6}$$

We therefore have that if $P(D|M_1)$ is computed within error ϵ and $P(D|M_2;E)$ is computed within error δ such that $\delta \leq P(D|M_2;E)/2$ and $\epsilon \leq P(D|M_1;E)/2$ then

$$\Delta\mathcal{B} = \left|\mathcal{B} - \frac{P(D|M_1;E)+\epsilon}{P(D|M_2;E)+\delta}\right| \leq \frac{|\epsilon| + 3\mathcal{B}|\delta|}{P(D|M_2;E)}. \tag{S7}$$

The overall error in the probability can arise from multiple sources but here we will focus on the error in the Bayes factors that arises from the use of a finite particle approximation to the probability density. Specifically,

$$P(D|H;E) = \int P(H)P(D|H;E)\mathrm{d}H \approx \tilde{P}(D|M;E) := \frac{1}{N_{\text{part}}}\left(\sum_{j=1}^{N_{\text{part}}} P(D|H_j;E)\right), \tag{S8}$$

for an ensemble of Hamiltonian models H_j drawn from the prior distribution $P(H)$. Here we use the convention that $\int f(H)\mathrm{d}H = \int f(H(\mathbf{x}))\mathrm{d}x^d$ where $\mathbf{x} \in \mathbb{R}^d$. We then have from Chebyshev's inequality that with probability greater than $1 - 1/k^2$ the error in this Monte-Carlo approximation is

$$\left|P(D|H;E) - \tilde{P}(D|H;E)\right| \leq k\sqrt{\frac{\int P(H)P^2(D|H;E)\mathrm{d}H - \left(\int P(H)P(D|H;E)\mathrm{d}H\right)^2}{N_{\text{part}}}} \tag{S9}$$

Now using the parameterization $H = \sum_{k=1}^{d} x_k \hat{h}_k$ for $x_k \in [-L, L]$. Similarly, let $P(D|M;E) = \int P(H)P(D|H;E)\mathrm{d}H$ we then have from the mean-value theorem that there exists a point in the domain of integration such that $P(\sum_k [\mathbf{y}(\mu)]_k \hat{h}_k) = P(D|M;E)$. we therefore have from Taylor's theorem that

$$P(D|H;E) \leq P(D|M;E) + \max \|\nabla P(D|\mathbf{x};E)\||\mathbf{x} - \mathbf{y}(\mu)|. \tag{S10}$$

Equations (S9) and (S10) then imply that

$$\left|P(D|M;E) - \tilde{P}(D|M;E)\right| \leq k\sqrt{\frac{P(D|M;E)\max\|\nabla P(D|H;E)\||\mathbf{x}-\mathbf{y}(\mu)|}{N_{\text{part}}}} \leq k\sqrt{\frac{P(D|M;E)\max\|\nabla P(D|H;E)\|dL}{N_{\text{part}}}} \tag{S11}$$

Thus we have that with probability at least $1-1/k^2$ that $\left|P(D|M;E) - \tilde{P}(D|M;E)\right| \leq \epsilon$ if

$$N_{\text{part}} \geq \frac{P(D|M;E)k^2\max\|\nabla P(D|H;E)\|dL}{\epsilon^2}. \tag{S12}$$

In order to make progress, let us assume that $\mathcal{B} \in \mathbb{R}^+ \setminus (1-\gamma, 1+\gamma)$. This assumption is needed in order to guarantee that there is a gap between the two possibilities. Thus if $\Delta\mathcal{B} \leq \gamma/2$ then the decision about the Bayes factor will not be qualitatively affected by the Monte-Carlo error due to finite particles. Eq. (S7) then tells us that it suffices to take

$$\epsilon \leq \frac{\gamma P(D|M_2;E)}{4}, \qquad \delta \leq \frac{\gamma P(D|M_2;E)}{12\mathcal{B}} \tag{S13}$$

which implies that if the first bound above dominates that it also suffices to take with probability at least $1-1/k^2$ the number of particles used to compute the first probability is

$$N_{\text{part}_1} \geq \frac{16P(D|M_1;E)k^2\max\|\nabla P(D|H_1;E)\|dL}{\gamma^2P^2(D|M_2;E)} \tag{S14}$$

Now let us further assume that for H_j in both model $M_j \in \{M_1, M_2\}$

$$\max\|\nabla P(D|H_j;E)\| \leq \kappa P(D|M_j;E). \tag{S15}$$

It then suffices to take

$$N_{\text{part}_1} \geq \frac{16\kappa\mathcal{B}^2k^2dL}{\gamma^2}. \tag{S16}$$

Similarly, we can choose

$$N_{\text{part}_2} \geq \frac{144\mathcal{B}^2P(D|M_2;E)k^2\max\|\nabla P(D|H_2;E)\|dL}{\gamma^2P^2(D|M_2;E)}, \tag{S17}$$

which is implied by

$$N_{\text{part}_2} \geq \frac{144\kappa\mathcal{B}^2k^2dL}{\gamma^2} \tag{S18}$$

Therefore if the number of particles given in (S18) is used for both models then a qualitatively accurate decision between the two models can be made according to the Bayes factors. The same result also holds if we only assume that $\mathcal{B} \in [1+\gamma, \infty)$.

This implies that Bayes factor calculation is stable under the SMC approximation provided that the gradients of the likelihood functions are not large (i.e. κdL is modest) and the promise gap γ is at least comparable to $\mathcal{B}$.

F. Scalability and Generalisation

The intrinsic scalability of the main *subroutines* of QMLA, i.e. BF analysis and QHL for parameter estimation, have already been discussed in § 1 C. Here we focus instead upon the implications of the graph search.

We first remind how all the $\mathcal{B}_{i,j}$ resulting from *consolidation* are assumed to be stored as a cDG representation across the model nodes. If all BFs composing the graph were computed using the same dataset, i.e. $D = \cup_{i,j\neq i} D_{ij}$, generating a DG would be immediately granted by the non-ambiguity of Eq. 1. This is not necessarily the case when $D_{ij} \neq D_{jk}$, which is most likely to happen if we run an independent instance of QHL per model. However, we expect the graph to converge to a DG as soon as enough statistical evidence is collected (i.e. once $\dim(D_{ij}) \gg 1\ \forall(i,j)$), because in this regime, differences across datasets become negligible. Therefore, we explicitly exclude cycles of ambiguous interpretation in the graph, by removing the edge which minimises $|\mathcal{B}-1|$, for each accidental cycle. Doing so and comparing systematically all model pairs leads naturally to the selection of a *layer champion*, corresponding to the node with highest indegree, without necessarily achieving the regime where BFs analyses generate naturally a DG.

The *exploration* stage poses the fundamental problem of how to explore sufficiently the space of possible models, without making it explode thus rendering QMLA pointless. A first question arises from the dimensionality of models entertained, i.e. the dimension of the corresponding Hilbert spaces: how should this be chosen?

One approach would be to explore only the dimension provided by a *dimensional witness* [13]. In this way, one could assume the dynamics of the quantum system to be entirely described via the special unitary group $SU(n)$, with n coincident with the lower–bound provided by the witness. Then, QMLA would be posed the task to learn which elements of the generator basis (e.g. the set Γ_n of generalised Gell-Mann complex $n\times n$ matrices γ_n) to include in the model, and with which parameters. That is, QMLA shall explore among subsets $\Gamma'_n \in \Gamma_n$ models of the form $\sum_{\hat{\gamma}_i\in\Gamma'_n} g_i\hat{\gamma}_i$ for the best Hamiltonian to reproduce measurements on the system. In this way, the new layer $\mu+1$ would simply e.g. include an additional γ_n from the set. The approach would thus share similarities with recent strategies characterising quantum *states* via neural networks [14–16]. In particular, when targeting an ideally black–box quantum device, also this strategy would employ implicitly a *dimensionality reduction* scheme, to map the (unknown) system interactions onto the possibly smaller size of the Hilbert space suggested by the witness.

Now, these kinds of mapping, if effective, are hard to be checked against intuition. As worries arise for the impact of artificial intelligence in the crisis of reproducibility affecting science [17], our intention is here to design an automated protocol bootstrapping the human characterisation process of a system, rather than replacing it. Therefore, we favoured exploring models constructed from a set of primitives easily interpretable: single–qubit Pauli operators $\hat{\sigma}_\alpha \in \Sigma_P$, i.e. $\alpha \in \{I,x,y,z\}$ with $\hat{\sigma}_I \equiv \hat{\mathbb{1}}$. In fact, it is known that any n–dimensional Hamiltonian can be written as the sum of tensor products of $\hat{\sigma}_\alpha$ [18]:

$$\hat{H}^* = \sum_{i,\alpha} h_i^\alpha \hat{\sigma}_i^\alpha + \sum_{i,j,\alpha,\beta} h_{ij}^{\alpha\beta}\left(\hat{\sigma}_i^\alpha \otimes \hat{\sigma}_j^\beta\right) + \ldots + \sum_{i,j,\ldots,n,\alpha,\beta,\ldots\nu} h_{ij\ldots n}^{\alpha\beta\ldots\nu}\left(\hat{\sigma}_i^\alpha \otimes \hat{\sigma}_j^\beta \otimes \ldots \otimes \hat{\sigma}_n^\nu\right) \tag{S19}$$

with Roman indices labelling the i–th subsystem on which operator $\hat{\sigma}_i$ acts. Most importantly, k–sparse, row–computable Hamiltonians, i.e. those Hamiltonians whose matrix has at most k non–zero elements per row or column (the locations of said elements being retrievable with an efficient classical algorithm), $\mathbb{H}_{\text{kcomp}}$ have been shown to admit a decomposition of the form in Eq. S19, with at most $\mathcal{O}(k^2)$ terms [19, 20]. The class $\mathbb{H}_{\text{kcomp}}$ is surprisingly wide, as it covers the cases of electronic Hamiltonians in quantum chemistry, as well as lattice model Hamiltonians (such as Ising and Heisenberg lattices [1, 21]). In all these cases, we can thus expect the depth of the *ideal* sDG to grow polynomially in k, *independently* from the size of the system n, an observation which is key in estimating the worst case scaling of QMLA.

Using Eq. S19 as a starting point to explore the Hilbert space, we are still left with the problem of establishing a ES to decide which terms to pick, and in which order.

For the purposes of this work, we addressed this question by generating layers that progressively increase the complexity of candidate models. In this way, we can improve the *interpretability* of the final DG, as deeper graphs correspond to searches that converged to models of higher complexity. In general, we assume n not known in advance, even if dimensional witnesses might bootstrap the guess for the root model(s) of the sDG. In such cases, the user can decide whether to assign a particular role to the dimension of candidate models, or not.

In a first implementation of QMLA, summarised in Fig. 2, we adopted a *greedy* ES designed to preferentially introduce models $\hat{H}_j$ in the *spawned* layer $(\mu+1)$, exploring the same Hilbert space dimension n_μ as the consolidated layer μ. Therefore, a lower dimensionality of the terms is privileged in the search [**Note1**], which first exhausts all *available* combinations of appropriate tensor products of the primitives $\hat{h}_i \in \Sigma_P$. Only once the latest μ has saturated the maximum number of independent operators allowed by n_μ (or by user-defined rules), QMLA will introduce terms such that the new layer involves a dimension $n_{\mu+1} = n_\mu + 1$.

Exploring models of lower dimensionality first can help the interpretability of the resulting DG, but is not necessary for QMLA to operate successfully. We demonstrated this with the analyses in Fig. 3, where we adopted a *genetic* ES instead. In this case, the search does not abide to choosing terms implying a lower-dimensional system first, but can populate the newly spawned layer with potentially any of the available terms.

Now, even if we admit the system Hamiltonian $\hat{H}_0 \in \mathbb{H}_{\text{kcomp}}$, the exploration phase is still affected by a major scalability issue: if the nodes of the ideal sDG scale polynomially, this is not true for any sDG, that explores brute–force also other potentially valid models. This is most easily seen by studying a specific case which covers many systems of interest: truncating the expansion in Eq. S19 to only include *pairwise interactions*, i.e. the first non–trivial terms. The number of all possible terms Θ that can be generated in this instance scales as $\mathcal{O}\binom{n}{2}$, i.e. the sDG depth grows combinatorially with the system's size. Additionally, in general also the maximum width of the DG grows combinatorially, as the generic layer μ includes models with any of $\mathcal{O}\binom{\Theta}{\mu}$ terms' combinations. This intractability is well known in the field of structure learning via graphical models, and the most direct way to deal with it is the introduction of a *greedy* exploration phase [22, 23]. That is, local optimisation is favoured against global optimisation, to keep the size of the global search DG manageable.

In QMLA, the greediness in the search is imposed via the *inheritance* rule: all models in layer $(\mu+1)$ expand upon the Hamiltonian form of their common parent node that is the champion node in μ (see also Fig. 1f). Inheritance rules can be seen as an extreme *pruning* rule for all non–champion nodes in μ (Fig. 1d), similarly to what adopted for a graphical model exploration in [12]. Inheritance might eventually require a dimensional matching when $n_{\mu+1} \neq n_\mu$: e.g. 1–qubit terms with the operator $\hat{\sigma}_z^1$ will be updated as $\hat{\sigma}_z^1 \otimes \hat{\sigma}_I^2$ when moving to 2–qubits candidate models. This greediness dramatically reduces the global number of models considered in the exploration. Adopting again the pairwise model, the overall number of nodes explored in the worst case is expected to be downsized:

$$\mathcal{O}\left[\prod_{\mu=1}^{\Theta}\binom{\Theta}{\mu}\right] \rightarrow \mathcal{O}\left[\sum_{\mu=1}^{\Theta}\binom{\Theta}{\mu}\right] \tag{S20}$$

when the inheritance rule is adopted. An interesting consequence of the inheritance rule is that the generic layer μ can then be interpreted as encompassing potentially all truncations of Eq. S19, that include exactly μ terms.

Nevertheless, the huge improvement exemplified in Eq. S20 is not enough to make QMLA tractable already for mid–sized systems, and highlights the discrepancy between the expected polynomial scaling of terms and a search through a combinatorially growing models' space. However, the inheritance rule becomes a powerful heuristic, when combined with limited aprioristic knowledge about the system to characterise. Continuing upon our example involving only pairwise interactions, the worst case scaling becomes quickly manageable by making two reasonable hypotheses: (i) that significant pairwise interactions involve nearest–neighbours, and (ii) that the index labelling is representative of neighbouring conditions. In this case, the first layer of the DG chooses but one out of n choices, and the overall models explored in the worst case is $\mathcal{O}\left[\sum_{\mu=1}^{n}(n-\mu+1)\right] = \mathcal{O}(n^2)$. One can use the last layer of the learnt DG to then progressively explore alternative hypotheses like index inversions, or longer–range interactions, without incurring straightforward in the intractable scaling suggested by Eq. S20.

Now that issues in the scaling of the sDG nodes have been discussed, we are left with the scaling of the number of edges in the cDG. If we were to compare each model $\hat{H}_i$ against all others explored, the cDG would be a *complete* graph, whereby the number of edges, each corresponding to a costly BF calculation, grows combinatorially with the sDG size [24]. This was avoided by allowing only nodes within a layer to form a *clique*, i.e. a complete subgraph of the cDG, whereas the only comparisons among different layers occur via their respective champions (see Fig. 1). In this way, all nodes other than the champion node for a given layer are essentially discarded from the learning process, as they stop being compared with any other competitive model. Additionally, we remind the *collapse* rule introduced in the main text, which reallocates the sub-tree rooted at $\hat{H}_{C(\mu)}$(if any) to $\hat{H}_{C(\lambda)}$ as the new parent node, discarding the whole layer of $\hat{H}_c$, effectively replacing inter-layer edges.

In conclusion, the efficiency in the QMLA's exploration of the models' space is implicitly a function of the a priori information provided, and how greedily the ESs are designed. towards local optimisation are the choices performed at each generation stage in the process. This trade–off resembles the limitations in the accuracy of the predictions that can be attained, versus the rapidity and human–interpretability of

the search that are typical of many machine learning methods, and in particular of graphical classification modelling [22, 25, 26]. In this regard, QMLA is designed to allow the user a degree of flexibility in managing such trade–off, increasing the explorative nature of the QMLA and hence its capability of approximating the global optimum, eventually at the expense of the overall cost of the procedure. For example, a greedy approach favouring local optimisation can be exacerbated or mitigated by:

1. Allowing or not collapse events. Terms provisionally inherited might prove superfluous once the model is expanded, but can be removed by inheriting directly from a grandparent node after a collapse event.
2. Replace the champion rule with a less greedy pruning rule. I.e. only those particularly unsuccessful models $\hat{H}_r$ are deactivated, that exhibit: $\mathcal{B}_{rj} < b \ , \ \forall j$ in the same layer, with b a user–defined threshold. All nodes kept active can then act as parents for additional spawned layers. This can be further generalised by
3. Using a genetic ES that allows mutations, because mutations naturally relax the necessity to retain an inherited term throughout the rest of the DG exploration, as the corresponding gene might be deactivated with a probability dictated by the genetic algorithm.

G. A comparison of QMLA and other characterisation methods

As mentioned in the main text, the problem of automating the characterisation of quantum systems and devices alike is far from new, and several efforts, in different directions, have been dedicated to the task. In this section, we intend to briefly overview some crucial novelty and/or differences of QMLA when compared to these methods. We will focus in particular upon tomographic and Hamiltonian reconstruction methodologies. A complete characterisation of a (sufficiently) isolated quantum system requires to know both its current state, as well as the dynamics induced by the system itself.

We shall skip commenting about quantum state tomography (QST) and refer to excellent reviews available to the reader [27, 28]. The reason is that the ultimate purpose of QMLA is to characterise quantum *processes*, instead of *states*, and therefore the two approaches solve radically different tasks. This is valid also for QST methods involving machine–learning approaches to reconstruct states [16, 29, 30], as well as methods invoking time evolution as a tool to discriminate among quantum states (e.g. Dynamical Quantum Tomography, [31]).

A closer approach to QMLA is instead provided by those tools intending to investigate the dynamics actually implemented by the device [32], going beyond the mere estimation of output states. This is the case e.g. of *Quantum process tomography* (QPT), currently the gold standard for system characterisation. QPT aims to provide an accurate reconstruction of the completely–positive linear map $\mathcal{M}(\rho_{\text{in}})$, describing a generic system's dynamics [9]: $\mathcal{M} = \sum_i \hat{A}_i \rho_{\text{in}} \hat{A}_i^\dagger$, with $\{\hat{A}_i\}$ a set of operators satisfying $\sum_i \hat{A}_i^\dagger \hat{A}_i \leq \hat{\mathbb{1}}$. A well–established classification distinguishes between *direct* and *indirect* QPT, whereby in the first case measurements are used to provide direct information about the underlying process, whereas indirect methods accomplish QPT via QST performed upon a set of probe states evolved according to $\mathcal{M}$ [33]. All methods proposed so far for exact QPT suffer from intrinsic scalability issues. Indirect methods inherit the lack of scalability present in QST, and require a total number of measurements scaling as $\mathcal{O}(d^1 6)$ [27], with d the dimension of the system Hilbert space. Ancilla–assisted process tomography trade an enlarged Hilbert space with a reduced number of total measurements, whereas methods bypassing QST tend to exhibit $\mathcal{O}(4^d)$ scaling and invoke single- and two- body interactions [33].

A first strategy to alleviate the gargantuan resources requested by QPT is to resort to approximate tomographic techniques. A highly successful proposal is based upon *compressive sensing*, which exponentially reduces the number of measurements required, by imposing the condition of sparsity of $\mathcal{M}$ [34]. These approaches alleviate the experimental efforts with the quantum device, but invoke non–scalable classical optimisation to perform the reconstruction of $\mathcal{M}$ from the compressed data, with the most recent studies addressing up to 5 qubits [35]. Very recently, the application of tensor networks to provide an efficient, compressed representation of $\mathcal{M}$, coupled with unsupervised machine learning techniques to perform the classical optimisation routine has ultimately allowed to expand the systems targetable via (approximate) QPT to 10 qubits.

A completely different approach addresses scalability issues via tomography–free methods, such as the *randomised benchmarking*. These focus on gate–implementations of the system dynamics, and replace accurate state preparation and measurement with fidelity measures and randomly chosen gates.

If all tomographic strategies have the advantage of assuming potentially no prior knowledge about the system, the information they extract into the map $\mathcal{M}$ assumes a non–accessible dynamics, if not for the output state. By doing so, these methods are not particularly apt to match those systems and devices, where there is a degree of control over the interactions among system parts, or with the environment. The duration of these interactions is particularly prone to accurate experimental control, as it is the case e.g. in superconducting as well as solid–state quantum architectures, whereby the implementation of gates rests upon accurate control of control pulses duration. Therefore, we argue how completely characterising a system at a single snapshot in time might not be the most suitable whenever it is possible to alter, in a controlled fashion, the system's evolution time, as discussed in the main text. A closer approach to QMLA is then offered by alternative methods attempting at inferring the Hamiltonian, and hence the generator of the dynamics, for the unknown system. By doing so, these methods allow a deeper understanding of the mechanisms underlying the observed dynamics which, in particular, translates into making available information apt to better *control* the device; whereby process tomography only retrieves a map capable of reproducing the system's behaviour. This category is composed by methods that retrieve the system's Hamiltonian, provided that its parametrisation is known. Using the same nomenclature of the main paper, the necessary a priori information is here a decomposition in a set of primitives $\{\hat{h}_i\}$, so that $\hat{H}_0 = \sum_i x_i \hat{h}_i$, and the characterisation task is to find optimal values for the parametrisation $\mathbf{x}$. This decomposition might derive from intuition about the system [2, 3], knowledge about its eigenstates [36], or be a generic basis of operators for the corresponding Hilbert space [37, 38], Methods belonging to this category thus provide possible implementations for the parameter learning phase of QMLA (see also Fig. 1b). A possibility is to perform this entire step with classical postprocessing, as it is the case of compressive Hamiltonian estimation (CHE, [37]), Eigenstate–to–Hamiltonian Construction (EHC, [36, 38]) or the same CLE introduced in the main text [2]. These methods typically suffer of scaling limitations, even when taking into account the sparsity of $\hat{H}_0$ to reduce the parameters to be estimated to a much smaller subset than the daunting 2^{2n} for an n–qubit system. For CHE, non-scalability occurs because of the complex optimisation problem invoked. EHC can retrieve a local $\hat{H}_0$ spanning a subsystem of volume $|L|$ with resources that scale linearly in $|L|$, at the cost of having experimental access to a number of trusted operations upon the evolved system state, which scales lienarly with the system size. This can be highly advantageous to recover the Hamiltonian governing a limited region of a target system, yet encounters an exponential wall in n, if the reconstruction of the global system Hamiltonian is attempted. In the case of CLE, the bottleneck is due to the classical exponentiation of the system's Hamiltonian. It is worth mentioning how CLE can be immediately generalised to a scalable approach, Interactive Quantum Likelihood Estimation (IQLE, [3, 4]), provided access to a quantum simulator, and a coherent channel capable of extracting ρ_{glo} as the input state of the quantum simulator (see § 1 C 2). The possibility to replace CLE, used here, with its efficient counterpart IQLE, was the main reason behind adopting QHL in this paper, to solve the parameter estimation task.

The possibility to reduce QMLA, bypassing the sDG construction, to simply estimate the most complex hypothetical parametrisation $\hat{H}_j$ deserves a discussion aside. Inferring which of the terms retrieved are actually significant would require imposing some threshold ϵ, such that all terms $\hat{h}_i$ whose corresponding $x_i < \epsilon$ would be neglected. However, choosing ϵ would be completely arbitrary instead of founded upon solid statistical evidence. For example, already in the case studied in the paper, the frequency parameters involved range from hundreds of kHz to a few GHz, posing severe difficulties in where to establish a meaningful threshold. It is easy to think how more extreme cases might lead to a wider span of these parameters, and the algorithm conclusion would completely rest upon an arbitrary choice of what to consider "significant enough". Parameter estimation applied to the most complex model would be inefficient and naturally prone to overfitting. This is because all methods quoted above solve the inverse problem of finding $\hat{H}_0$ by undergoing a (non–)convex optimisation. In the case of CLE, it was shown repeatedly how the more parameters we involve, the longer the training phase for the model becomes, at the expense of the efficiency of the procedure. Most importantly, including several non–significant parameters can lead to severe overfitting, with poor performances at reproducing the dynamics in test–datasets which are much different from the training–dataset. A severe such case is reported in Fig. S7, where the sDG was allowed to explore depths much further than what supported by statistical evidence. Finally, progressing towards more complex models rests upon the choice of the user, and is hence far from optimal, nor automated. We can easily exemplify this with one of the study cases reported in the main paper, the long decay study in Fig. 3. Using the naive guess available about the system [39], one might attempt learning parameters of a Hamiltonian with hundreds of qubits, corresponding to the ^{13}C nuclei potentially coupled to the NV centre , a task beyond the capabilities of our classical simulators. Even assuming such computational power available, the aforementioned inefficiencies

would arise in QHL methods, or non–uniqueness issues in EHC. On the contrary, the user might oversimplify by mapping the problem onto a space much smaller than needed, e.g. including $n^s < 10$ spins. Both in QHL and in EHC, this would manifest by non–convergence in the learning of the parameters, but this would not inform the user about how to improve the model.

Focusing on NV centres, this latter limitation applies to all empirical approaches used to infer information about the environment surrounding the NV electron spin. Early approaches inferring coupling parameters from echo modulation [39] up to the most recent Bayesian techniques to extract actionable information about the nuclear spin bath [40]. All assume a model for the system whose parameters are learned, with terms fixed a priori, without including the possibility to modify it according to observations.

Finally, we acknowledge how the flexibility in the search phase of QMLA leading already in this paper to two possible variants, might be susceptible to further modifications, e.g. according to the specific problem addressed. For example, we have observed how a completely uninformed search can be best performed by a probabilistic, unstructured GR, when across a model space including instances with a specific pattern (as the cluster expansion invoked to solve the case illustrated in Fig. 2 of the main text). Here, we have built such a ES with a genetic algorithm, mapping features of each possible expansion on genes. However, other possibilities are also viable, such as Orthogonal Matching Pursuit, also used for feature selection, and specifically to refine GA-generated solutions in the specific realm of retrieving energetically–optimal lattice configurations [41]. Clearly, the adoption of any exploration strategy implemented by the GR shall occur bearing in mind their different merits.

In conclusion, more than offering an alternative to the methods listed above, QMLA encompasses them, and can leverage upon the advantages of each to implement the key subroutine of Hamiltonian parameters' estimation.

2. Details about the implementation of QMLA

A. QMLA implemented with a classical simulator

We test QMLA using classical likelihood estimation (CLE) for parameter estimation, and adopting the *Phase 3* of *Blue Crystal*, the supercomputer of the University of Bristol, composed of 223 nodes each made of 16 × 2.6 GHz SandyBridge cores as the classical simulator. Adopting a supercomputer is justified by the need of running a large number of independent instances of the protocol, each involving up to 6 qubits. These simulations retain full control over the correct model form and parameters, and no experimental limitation due to setup performances, input probe states and noise (other than numerical approximations available on a 64-bit architecture). Simulations are coded in `Python` through the QMLA framework, adopting the `Numpy, Scipy, Qutip` and `Qinfer` libraries.

Our implementation allows for QMLA instances as well as single QHL instances to run in parallel, and likewise BF calculations. The most expensive operation is the calculation of likelihoods $|\langle d|\, e^{-i\hat{H}'t}\, |\psi\rangle\,|^2$, due to the exponentiation of $\hat{H}'$ scaling exponentially with the system's size. Note how a (trusted) quantum simulator is expected to compute these quantities in polynomial time, making a transition to QLE a necessary step to circumvent this intrinsic limitation of classical simulators. As an example, 2-qubit Hamiltonians are exponentiated typically in $t_H \sim 0.5$ ms, using `scipy.linalg.expm`. QHL involves sampling of N_P particles in a single epoch, over N_E epochs, so each QHL instance requires $N_P N_E t_H$. Pairwise BF calculations involve a further N_e epochs on each model in the pair, i.e. a cost of $2N_P N_E t_H$, for all the combinations $N_{BF}(\mu)\binom{N_m(\mu)}{2}$ of the $N_m(\mu)$ nodes within μ. After all layer champions are determined, a parental collapse rule involves pairwise BF between parent/child models, i.e. up to $N_\mu - 1$ BFs calculations, with N_μ the depth of the DG. These contributions lead to a total runtime for a single QMLA instance:

$$T \sim t_H \times \left(\sum_\mu \left\lceil \frac{N_m(\mu)}{p} \right\rceil N_P N_E \; + \sum_\mu \left\lceil \frac{N_{BF}(\mu)}{p} \right\rceil 2 N_P N_E \; + 2 \left\lceil \frac{N_\mu - 1}{p} \right\rceil N_P N_E \; + 2 \left\lceil \frac{N_C}{p} \right\rceil N_P N_E \right) \quad \text{(S21)}$$

when leveraging upon p available processes. The QMLA instances reported in Fig. 1 have $T \sim 20hrs$, resulting from $t_H \leq 0.5$ ms, $N_\mu = 9, N_E = 1000, N_P = 3000$ and used $p = 6$ parallel processes. Finally, $N_m(\mu)$ can be inferred from the DG in Fig. 2f. Increasing N_P, N_E typically improves the learned parameterisations for each model, at the price of increasing the computational time required.

We mentioned how QHL methods encode the (a priori) information about the parameters in a prior distribution, $P(\mathbf{x})$. Preliminary information about the expected order of magnitude of the parameters involved in the model Hamiltonians can then be used to make QHL stable without an excessive number of particles (see § 1 E), and correspondingly enhance the computational performance of the protocol. For all instances studied here, we thus chose initial priors which are informative of said order of magnitude, m. In particular, we start with normal priors $5 \pm 2MHz$ for all parameters.

B. Results

1. *Parameter Inference*

Fig. S2a&b show the values of the parameters learned by each champion model for cases (*ii*) "simulated" and (*iii*) "experimental" respectively, described in the main paper at pg. 3. For reference, we report the true values of the (applicable) parameters, for the simulated case. These histograms make evident how many champion models contain each term reported.

The occurrence for each Hamiltonian term, along with the confidence in parameter values, allow us to draw conclusions about the physics of the NV centre electron spin. In particular, we find that the element of the hyperfine tensor $\hat{A}_z \equiv \hat{A}_\parallel \sim 2$ MHz, a finding in good agreement with theoretical expectations [42]. Interestingly, in few cases QMLA favours models which might be interpreted as overfitting, in the light of the values of corresponding parameters. For instance, in the simulated case of Fig. S2a, terms $\hat{A}_x, \hat{A}_y$ are preserved in 22 and 29 champion models respectively, even if their corresponding parameters have median $|x| < 0.1$, suggesting how these terms, which are neglected in usual approximations (see § 3 A), might still have a residual impact over the system's dynamics.

$\dim(\hat{H})$ (qubits)	Operator form	Abbreviation
1	$\hat{S}_\alpha$	$\hat{S}_\alpha$
1	$\sum_{i\in\{\alpha,\beta\}} \hat{S}_i$	$\hat{S}_{\alpha,\beta}$
1	$\sum_{i\in\{x,y,z\}} \hat{S}_i$	$\hat{S}_{x,y,z}$
2	$\sum_{i\in\{x,y,z\}} \left(\hat{S}_i \otimes \hat{\mathbb{1}}\right) + \hat{S}_\alpha \otimes \hat{A}_\alpha$	$S_{x,y,z}\hat{A}_\alpha$
2	$\sum_{i\in\{x,y,z\}} \left(\hat{S}_i \otimes \hat{\mathbb{1}}\right) + \sum_{i\in\{\alpha,\beta\}} \left(\hat{S}_i \otimes \hat{A}_i\right)$	$S_{x,y,z}\hat{A}_{\alpha,\beta}$
2	$\sum_{i\in\{x,y,z\}} \left(\hat{S}_i \otimes \hat{\mathbb{1}} + \hat{S}_i \otimes \hat{A}_i\right)$	$S_{x,y,z}\hat{A}_{x,y,z}$
2	$\sum_{i\in\{x,y,z\}} \left(\hat{S}_i \otimes \hat{\mathbb{1}} + \hat{S}_i \otimes \hat{A}_i\right) + \hat{S}_\alpha \otimes \hat{A}_\beta$	$S_{x,y,z}\hat{A}_{x,y,z}\hat{T}_{\alpha\beta}$
2	$\sum_{i\in\{x,y,z\}} \left(\hat{S}_i \otimes \hat{\mathbb{1}} + \hat{S}_i \otimes \hat{A}_i\right) + \sum_{ij\in\{\alpha\beta,\gamma\delta\}} \left(\hat{S}_i \otimes \hat{A}_j\right)$	$S_{x,y,z}\hat{A}_{x,y,z}\hat{T}_{\alpha\beta,\gamma\delta}$
2	$\sum_{i\in\{x,y,z\}} \left(\hat{S}_i \otimes \hat{\mathbb{1}} + \hat{S}_i \otimes \hat{A}_i\right) + \sum_{ij\in\{xy,xz,yz\}} \left(\hat{S}_i \otimes \hat{A}_j\right)$	$S_{x,y,z}\hat{A}_{x,y,z}\hat{T}_{xy,xz,yz}$

Table S1 List of all models used in the QMLA implemented in Fig. 1, along with the corresponding number of qubits and the abbreviation used in the text. Single terms adopt the same notation as in Eq. 3b. Parameters are omitted for brevity.

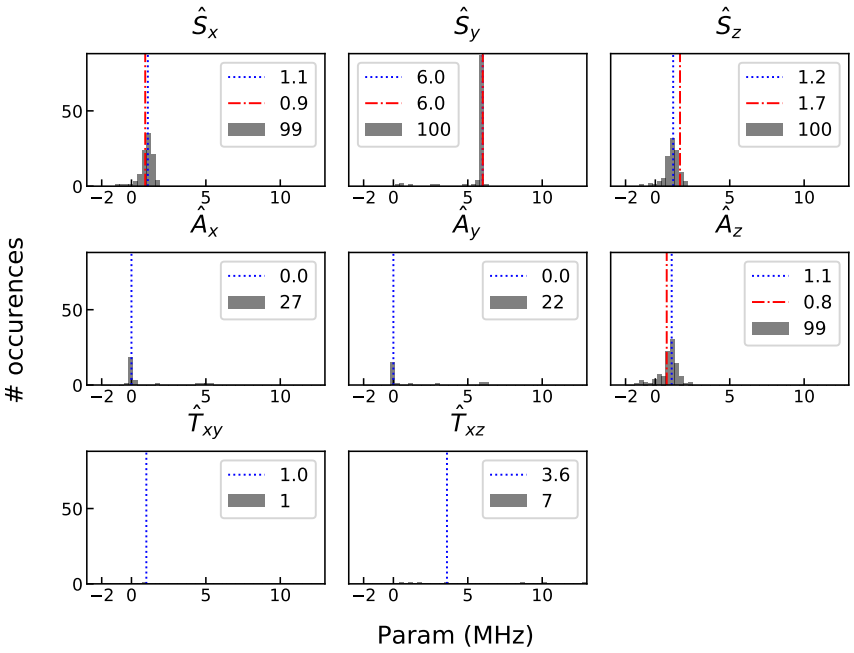
2. *Dynamics Reproducibility*

In order to capture the capability of a given model $\hat{H}_j$ to reproduce the dynamics generated from $\hat{H}_0$, we adopted the coefficient of determination R^2 as a figure of merit, computed against an equally spaced grid of datapoints up to $t = 5\mu s$ for simulated cases (*ii*, see the main text at pg. 3), or against the whole available dataset for experimental cases (*iii*). In Table S2, we list the number of QMLA instances won by each model, among those selected as $\hat{H}'$at least once. We also report the corresponding R^2. A crucial observation is that typically those models exhibiting the best agreement with data drawn from $\hat{H}_0$ (and hence a higher R^2) tend to win more frequently QMLA instances (as it is expected). More interestingly, BFs operate strong preferences towards simpler models, when the predictive power estimated by R^2 is similar.

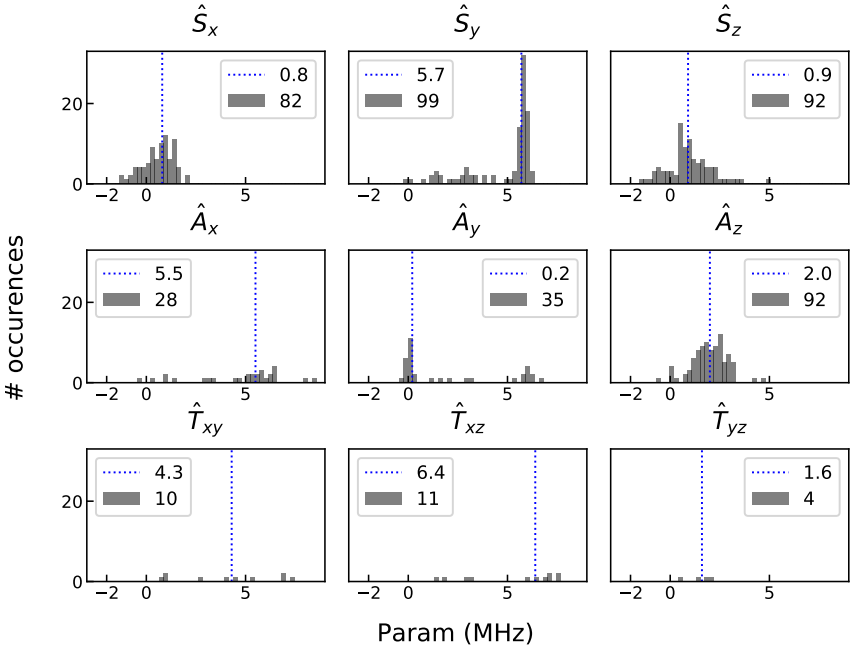
Finally, to better visualise the agreement between experimental (or simulated, target) data and the predictions operated by the same models listed in Table S2, we present in Fig. S3 the plots for the likelihoods attained in both cases. It is evident how, in most cases, the nominated champions models can closely emulate the system to characterise, and this is particularly true for times before a first decay in the signal occurs (see also § 2 C for further discussion).

Model	Experiment		Simulation	
	Wins	R^2	Wins	R^2
$\hat{S}_{y,z}\hat{A}_z$	9	0.8	1	0.26
$\hat{S}_y\hat{A}_{x,z}$	2	0.63		
$\hat{S}_{x,y,z}\hat{A}_z$	45	0.86	61	0.97
$\hat{S}_{x,y,z}\hat{A}_y$			1	-0.54
$\hat{S}_{x,y,z}\hat{A}_{x,y}$	3	0.81		
$\hat{S}_{x,y,z}\hat{A}_{y,z}$	14	0.83	10	0.96
$\hat{S}_{x,y,z}\hat{A}_{x,z}$	6	0.64	15	0.99
$\hat{S}_{x,y,z}\hat{A}_{x,y,z}$	2	0.72	5	0.97
$\hat{S}_{x,y,z}\hat{A}_{x,z}\hat{T}_{xz}$			1	0.68
$\hat{S}_{x,y,z}\hat{A}_{x,y,z}\hat{T}_{xz}$			5	0.77
$\hat{S}_y\hat{A}_{x,y,z}\hat{T}_{xy,xz,yz}$	2	0.31		
$\hat{S}_{x,y,z}\hat{A}_{x,y,z}\hat{T}_{xy,xz}$	4	0.67	1	0.32

Table S2 QMLA results for experimental data and simulations. In simulations, $\hat{S}_{x,y,z}\hat{A}_z$ was chosen as the *correct* model. We state the number of QMLA instances won by each model and the average R^2 for those instances as an indication of the predictive power of winning models.

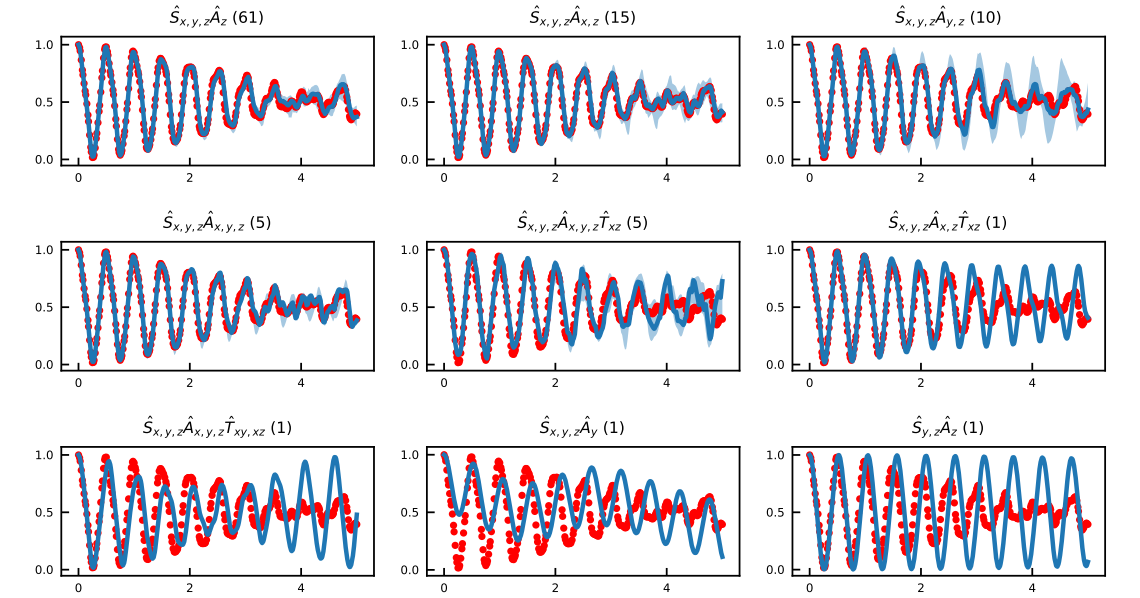


a, Simluated QMLA instances. Red dotted lines show the true parameters.

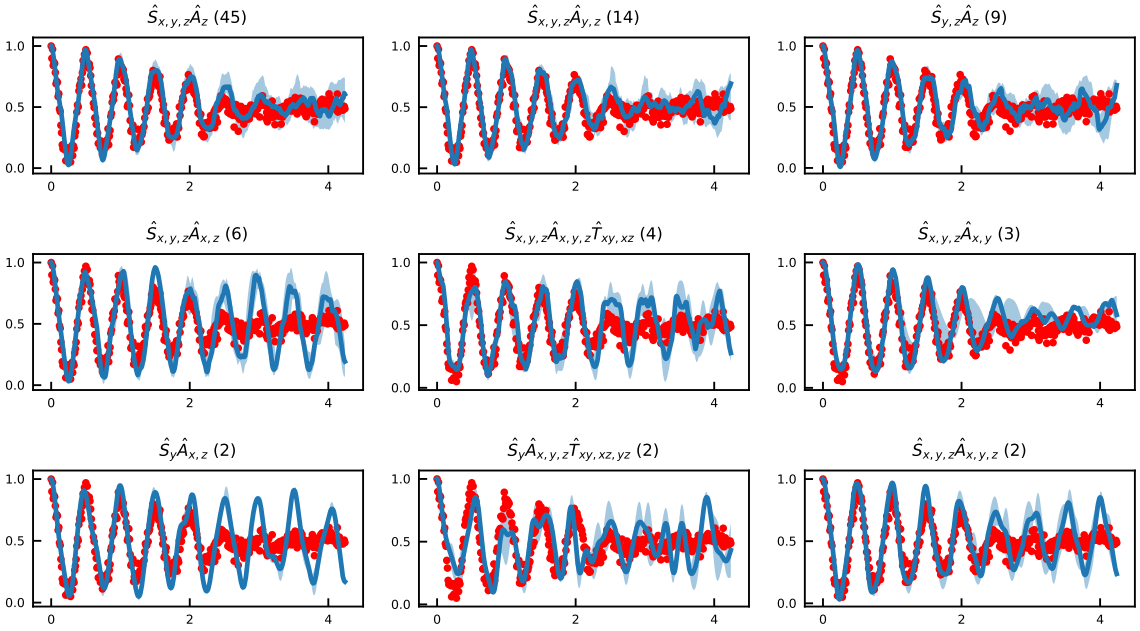


b, Experimental QMLA instances.

Figure S2 Histograms for parameters learned by champion models on (**a**) simulated and (**b**) experimental data. Blue dotted lines indicate the median for that parameter; grey blocks show the number of models which found the parameter to have that value and the number listed in the legend reports how many champion models contained that term.



a, Simulated data



b, Experimental data

Figure S3 Dynamics reproduced by various champion models for simulation and experiment. Expectation values are shown on y–axis with time on x–axis. Red dots give the true dynamics of $\hat{H}_0$, while the blue lines show the reconstruction by $\hat{H}'$. $\hat{H}'$ is listed on top of each plot, with the number of QMLA instances that model won in brackets.

C. Finite-size effects of the nuclear spins bath and reversibility of the dynamics

A well-established modelling of decoherence effects in open quantum systems involves the introduction of *indirect measurements* occurring on the quantum system, when operations occur that effectively trace

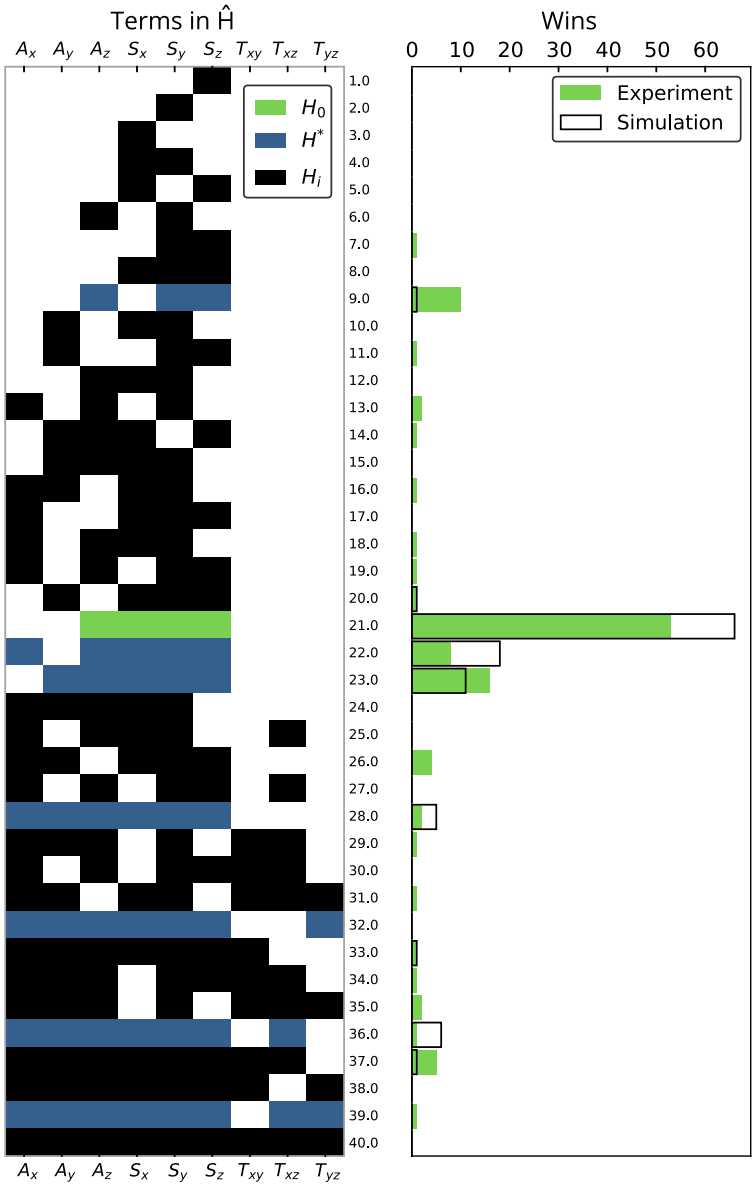


Figure S4 Left: map of the various Hamiltonian terms included in each of the possible 40 candidate models explored by QMLA, in instances whose analysis was reported in Fig. 2 of the main text. The explicit form of each operator can be inferred from Table S1. IDs of candidate models are on the vertical axis, and labels for the terms on the horizontal one. The true model $\hat{H}_0$ for the simulated case is highlighted in green and the subset of credible models which were generated by the ES in blue (see § 3 A). Right: wins for each of the candidate models as above out of 100 independent QMLA runs, reported as a histogram for cases adopting simulated data (empty bars) or the experimental dataset of Fig. 2e (green bars).

out environmental degrees of freedom, which have become correlated with the system state [9]. In the main paper we have systematically considered models inclusive of Hamiltonian terms of the form $\hat{S} \cdot \hat{I}_\chi$ (see Eq. 3a), which is a standard interaction term leading to system-bath correlations [9, 43].

In the theoretical framework of open quantum systems, it is customary to model decoherence phenomena introducing a bath of n^s spins [43], or harmonic oscillators [44]. The reason to adopt a bath of interacting systems is to ensure an irreversible decoherence, as usually observed in real physical systems. This adaptation can be seen by using Poincaré recurrence theorem. Assuming the state of the global quantum system starts as a separable state $|\psi(t=0)\rangle = |\psi\rangle_{\text{sys}} \otimes |\phi\rangle_{\text{env}}$, the theorem states that a time $\tilde{t}$ exists, such that the evolved

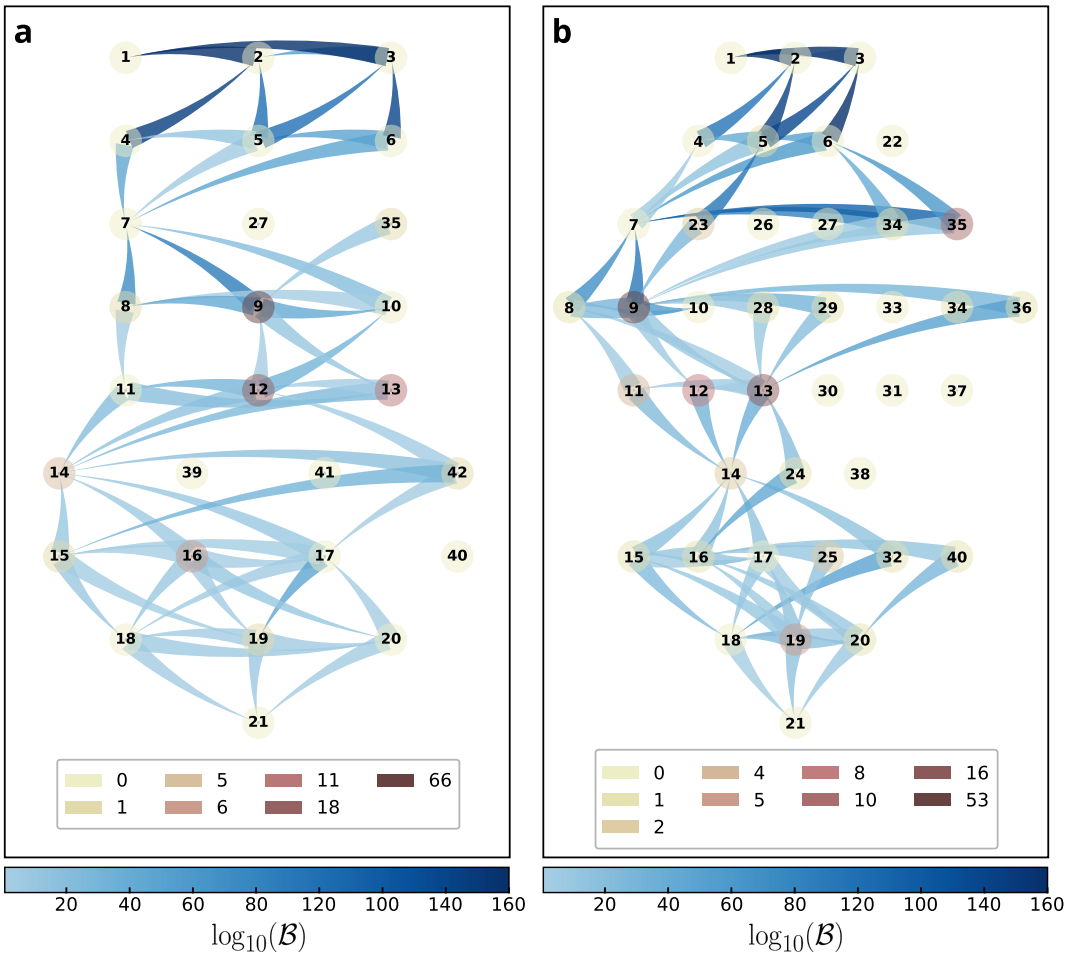


Figure S5 **a**: cumulative sDG obtained out of 100 independent QMLA instances, for the same set of simulations reported in Fig. S2a & S3a. All edges reported have been estimated at least once during the protocol. For each edge, we compute the median value of the corresponding Bayes factor ($\mathcal{B}$) out of all the sDG instances including an edge among the same nodes. The resulting median BF is reported colour–coded, as in the colour–bar scale at the bottom. Within the boxed legend, we associate different colours to those nodes (i.e. models) winning at least a QMLA instance. The corresponding number is also elicited. **b**: same as (a), but for the QMLA runs agains experimental datasets, corresponding to Fig. S2b & S3b. Model labels are after S4. The information about the frequency, with which each edge is evaluated, was omitted from these plots, to render each edge clearly distinguishable. Note how a higher number of collapse events in (a) leads to additional models being explored.

state $|\langle\psi(0)|\psi(\tilde{t})\rangle|^2 \sim 1$, provided that the global system has a finite eigenspectrum of energies $\{E_n\}$. An immediate consequence is that at times $\mathcal{O}(\tilde{t})$ (known as Poincaré recurrence times [9]), the global state can become separable again, thus leading to *revivals* in the coherence of the system state. If we consider the simplest possible interaction term:

$$\sum_b |b\rangle_{\text{sys}} \langle b|_{\text{sys}} \otimes \hat{B}_{\text{env}}, \tag{S22}$$

with $\hat{B}$ a generic operator acting on the bath and $\{|b\rangle\}$ an orthonormal basis for the system, then the global evolved state can be represented as:

$$|\psi(t)\rangle = \sum_b p_b |b\rangle_{\text{sys}} \otimes |\phi_b(t)\rangle_{\text{env}} \tag{S23}$$

and therefore the decoherence can be characterised in terms of the overlaps $M_{b,b'}(t) = \langle\phi_b(t)|\phi_{b'}(t)\rangle$ [9]. Now, the decay of these overlaps, and their recurrence times, depend on the size of the bath, as a richer

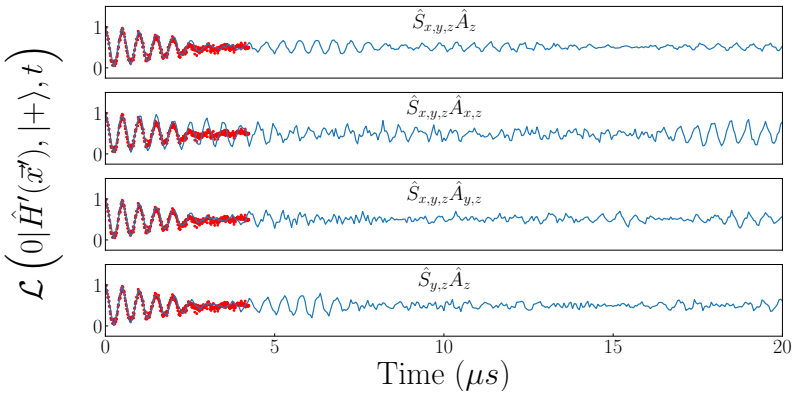


Figure S6 In solid lines, the predicted likelihoods $\mathcal{L}(0|\hat{H}'(\mathbf{x}'),|{+}{+}'\rangle,t)$ for different champion models selected by QMLA (as reported within each plot) for times exceeding the longest t experimentally available. The experimental dataset is reported for comparison as red dots in all subplots. QMLA was run with a greedy ES, mapping all environmental interactions upon a single additional qubit, as in Fig. 2 in the main text.

dynamics corresponding to a higher-dimensional bath will tend to increase the recurrence times at which $M_{b,b'}(t)$ returns non-negligible. In real systems it is empirically observed that a decay $M_{b,b'}(t) \propto \exp[-\Gamma_{bb'}t]$ occurs, which is in agreement with predictions from Markovian assumptions [45]. Generally, it can be proven that the recurrence times $\tilde{t}$ for the global system scale combinatorially with the bath size n^{s} [43]. Hence the necessity in simulations to increase the bath size for a system with a discrete spectrum, not to observe revivals as an artefact of the finite size of the bath. This behaviour is equivalent to what known for artificial quantum simulators, characterised by *finite-size effects* in the bath degrees of freedom (e.g. ion trap simulators [46]).

In the main paper (analyses summarised in Fig. 3), we have shown how the size of the bath can be inferred from *Hahn-echo* experiments. Instead of adopting an appropriate size, one might map all interactions with the environment upon a single qubit (as we did in the first part of the paper). As reminded, this could be due e.g. to match the maximum bath dimension attainable with the available simulator. However, this simplification leads inevitably to finite-size effects in the system dynamics. The learning procedure might parameterise opportunely candidate models $\hat{H}_j$ of reduced size to reproduce well short–decays (e.g. those reported in Fig. 1e). However, discrepancies due to the short Poincaré recurrence times will become observable in the predicted dynamics. This is shown in Fig. S6 for time ranges close to the short–decay, and in Fig. S7 for the longer times when the revivals from Hahn–echo occur.

Finally, we briefly remark how the standard way to address these artefacts –without introducing additional qubits– is the introduction of phenomenological decay terms $\exp[-\Gamma(t)]$ in the model Hamiltonian [2, 10]. We chose not to implement any such strategy in this work, and include only time–independent terms among the available primitives. The reason behind is that we favoured the learning of open system features from first–principles, via environmental qubits, instead of imposing a phenomenological modelling within $\hat{H}$.

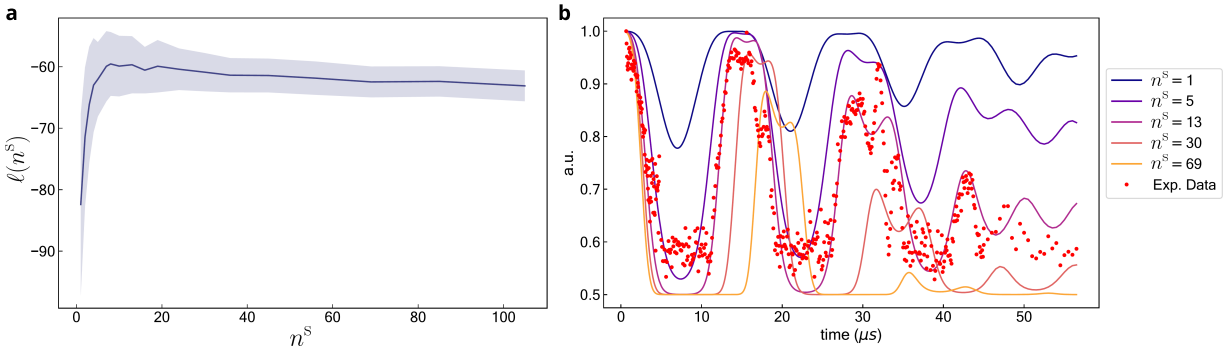


Figure S7 Further analysis of the hyperparametrised model (see Eq. 5). **a** Estimated log-likelihood for different tentative numbers n^s of environmental spins. The latter is progressively increased and, for each n^s, the values of the hyperparameters $\{\mathbf{B}_0, \mathbf{B}_1, \sigma_B, \omega_0, \delta_\omega, \sigma_\omega\}$ are learned via CLE, using 1000 particles along 100 epochs. Then, $\ell(n^s)$ is estimated for the cumulative training set. The procedure is repeated 150 times, the average $\bar{\ell}(n^s)$ is reported on the plot as a solid line, along with the 68.27% confidence interval as a shaded area. **b** Plots for the expected likelihood $\mathcal{L}(0|t)$, along with experimental data reporting the normalised detected PL. $\mathcal{L}(0|t)$ is reported colour-coded for models adopting different n^s, each with the averaged hyperparameters learned in **a**.

3. Experimental Details

A. Experimental system and sample description

The measurements were performed on a home-build confocal setup at room-temperature. In our setup we align an external magnetic field of $11mT$ parallel to the NV centre's axis, using Helmholtz-coils. This magnetic field lifts the degeneracy of the electron fine states $m = \pm 1$. For the experiments shown here we only use the $m = -1$ electron state. We perform Hahn-echo measurements by using off-resonant laser excitation ($532nm$) for initialisation and readout of the electron spin. To manipulate the electron spin in our Hahn-echo experiments, we apply a $\pi/2 - \pi - \pi/2$ MW-pulse at the resonant frequency, with varying time between the MW pulses. We use a strip line to apply microwave signals of a few microwatts. The pulse sequence is repeated about 3×10^6 times for each microwave frequency to acquire statistics, and the detected photon detection times are histogrammed leading to about 300 counts in each $25ns$ bin during the laser pulse. Single photons are collected by the confocal setup and detected by an avalanche photo diode. The detected counts at the end of each Hahn-sequence are normalised with respect to the mixed state of the initialisation pulse.

For the measurement we use an electronic grade diamond sample [110], with a 4 ppb nitrogen impurity concentration and a natural abundance ($\approx 1.1\%$) of ^{13}C (Element6).

Finally, in the main text we have outlined how the Hahn-echo experiment reported here can be modelled as the Hamiltonian in Eq. 3b, embedding hyperfine interaction terms between the NV centre electron spin system, and its surrounding environment of nuclear spins. Here we report a few remarks concerning the significance of the various terms in $\hat{H}_{\text{full}}$:

- The electron spin splitting $\Delta \sim 2.87$ GHz is large enough, to justify neglecting terms $\propto \hat{S}_x, \hat{S}_y$ from the model, as long as the external magnetic field is sufficiently aligned with the NV symmetry axis [6, 47];
- For the same reason, among off–diagonal terms in the hyperfine tensor $\mathbf{A}$, we expect the most significant to involve the NV symmetry axis z (hence those corresponding to $\hat{T}_{xz}$, $\hat{T}_{yz}$);
- The order of magnitude for the magnetic field can be inferred immediately from the approximate period of the observed Hahn-echo oscillations, which equates the bare Larmor frequency $\omega_0 = \mu_n g_n B = \gamma_n B \sim 66$ kHz for the data in Fig. 3f, confirming the interaction with ^{13}C atoms;
- Therefore, the contribution for the site occupied by the neighbouring nitrogen nucleus, which instead precesses with $\mathcal{O}$(GHz) frequency, plays a minor role in the observed dynamics (because of the usual rotating wave approximation);

- The factor for the electron spin precession reads $\mu_e g_e B_z/\hbar \sim 1.94$ GHz;
- **A** can be approximated as diagonal within the point–dipole approximation (and hence off diagonal terms $\hat{T}_{kl}$ should be weaker than their diagonal counterparts);
- Finally, the diagonal elements of the hyperfine tensor $A_{j,\parallel}$ and $A_{j,\perp}$ depend on both the atomic species at site j, as well as its distance from the electron spin [39], but are most commonly found to be in the order of hundreds of kHz to few MHz [42]; moreover, the terms $\hat{A}_y, \hat{A}_y$ (in the main text notation) can also be neglected within the 0^{th} order of the Floquet theory, i.e. the *secular approximation* [6, 47, 48].

Due to the considerations above, when running QMLA against data from the actual NV centre , it is not immediate to conclude for a "correct" model, as far as it invokes terms within Eq. 3b. This justifies the introduction of a set of credible models: $\{ \{\hat{S}_z\hat{A}_z\}, \{\hat{S}_z\hat{A}_z\hat{T}_{xz}, \hat{S}_z\hat{A}_z\hat{T}_{yz}\}, \{\hat{S}_z\hat{A}_z\hat{T}_{xz,yz}\}, \{\hat{S}_{xz}\hat{A}_{xz}, \hat{S}_{yz}\hat{A}_{yz}\}, \{\hat{S}_{xyz}\hat{A}_{xyz}\}, \{\hat{S}_{xz}\hat{A}_z, \hat{S}_{yz}\hat{A}_z, \hat{S}_{xyz}\hat{A}_{xz}, \hat{S}_{xyz}\hat{A}_{yz}, \hat{S}_{xyz}\hat{A}_z\}, \{\hat{S}_{xyz}\hat{A}_{xyz}\hat{T}_{xz}, \hat{S}_{xyz}\hat{A}_{xyz}\hat{T}_{yz}\}, \{\hat{S}_{xyz}\hat{A}_{xyz}\hat{T}_{xz,yz}\} \}$. Each of the groups corresponds to invoking one or more among the approximation quoted in this paragraph. Success rates for all of them can be inferred from Fig. S4 In particular, it is interesting to observe how the most successful model invoked by QMLA deems the simplest approximation for the hyperfine interaction (i.e. assuming only $A_{j,\parallel}$ to be non–zero) to provide a sufficient description. At the same time, QMLA consistently finds a non–negligible residual off–axis component of the external magnetic field **B**, as terms $\hat{S}_x, \hat{S}_y$ are kept.

B. Noise in the experimental setup

Bayesian-methods are known to exhibit intrinsic properties of noise robustness against i.i.d. and Markovian sources of noise [2, 49]. QMLA partially inherits such noise resilience by adopting QHL as a subroutine for parameter estimation and Bayes Factors (BF)s for model comparison. However, the role of noise in our protocol has more profound implications. Learning a model for the noise affecting the system is inextricable from learning a model for the system dynamics itself, in all those cases where the contribution from noise sources cannot be neglected [10, 50]. Therefore, we expect QMLA to attempt modelling noise sources within the framework of the primitives provided in the user–defined library. Apparent overfitting might then be interpreted, in some instances, as triggered by features in the dynamics not predicted by a noiseless model.

In the specific case of the NV centre electron spin studied in this paper, the hyperfine interaction with the nuclear spins in the bath can be in itself interpreted as a "noise" in the isolated system picture. However, its effects are so crucial that we deem necessary to expand the picture correspondingly. On the contrary, other, minor noise effects might not justify the attempt to reproduce them accurately by increasing the complexity of the model. The adoption of BFs is intended to guard against such unwanted overfitting tendencies of the QMLA, providing a solid statistical ground for deciding systematically in favour, or against, increases in the model complexity [11]. Nevertheless, we report here a review of noises that might lead to a spurious response in Hahn-echo experiments with NV centres, even to the point of invalidating the correct model $\hat{H}_{\text{diag}}$.

Noise in the active electron spin control. The Hahn-echo experiments in this paper require the implementation of three MW pulses (see also Fig. 2a in the main text), controlling respectively: i) the initial preparation of the electron spin in the $|+\rangle$ state, i.e. a $\pi/2$-pulse, ii) the Hahn rotation, i.e. a π-pulse, iii) the preparation of the evolved state for projection onto the measurement basis $\{|0\rangle, |1\rangle\}$, i.e. another $\pi/2$-pulse, with all rotations along the y-axis of the Bloch sphere representing the electron spin. Constant offsets as well as i.i.d. noise might affect all such operations. Accurate characterisation of the setup effectively minimises any offset affecting these operations, whereby the initial state can be represented as:

$$|\psi\rangle_{\text{sys}} = (|+\rangle + \omega\,|\psi\rangle)/\sqrt{1+\omega^2}, \tag{S24}$$

with $|\psi\rangle$ a normalised one-qubit state and $\omega \ll 1$ a single fixed parameter representing the severity of the offset. Residual offsets in the preparation and projection operations might reduce the visibility of the first peak, or slightly alter its occurrence in time. We partially counteract these effects by rescaling the observed PL (proportional to the likelihood $\mathcal{L}(0|\hat{H}, t, |\psi\rangle)$) to the full range $[0, 1]$, and adjusting the zero of the experimental time against the first peak in $\mathcal{L}(0|\hat{H}, t, |\psi\rangle)$. Moreover, to prevent QMLA from using recursively a biased probe state $|+\rangle$ not matching the experimental $|\psi\rangle_{\text{sys}}$, we randomise the probe prepared in the simulator adopting Eq. S24, with ω a random variable normally distributed according to $\mathcal{N}(0, \sigma_\omega)$, with $\sigma_\omega \sim 0.03$ as extracted from the system characterisation. Thus, the effect of a constant offset on the system

preparation is mapped onto a less disruptive randomised noise, easier to address via Bayesian-inference schemes. Similar randomisations might also be introduced in the simulator for the final rotation ahead of the projection measurement, and in the Hahn rotation angle, however these were deemed unnecessary in the light of the good agreement achieved by simulations. Offsets in the Hahn-angle implemented might instead screen less effectively the electron spin from bath effects. However, this should effect quantitatively (i.e. the strength and components of the hyperfine interaction) more than qualitatively the results in the learning (i.e. the form of the model).

Optical readout noise. This source of noise originates from the optical readout, i.e. the $\sim 33\%$ contrast in PL detected as the final electron spin state is in the $|0\rangle$ (bright) as opposed to the $|1\rangle$ (dark) stable states. The normalised PL photon counts $C_0(t,t')$ from the experiment are proportional to $\mathcal{L}(0|\hat{H},|\psi\rangle,t,t')$, which is obtained by subtracting the dark counts (i.e. the floor of the C_0 signal) and mapping the effective number of surplus counts onto the full interval $[0,1]$. Assuming that the noise floor counts are approximately constant, then C_0 is affected by two sources of noise: losses due to imperfect collection efficiency, and Poissonian-noise in the counts readout. Losses are known to impact negatively the learning capability of Bayesian-inference schemes in the absence of error modelling, when the number of counts is extremely low, whereas can be safely neglected already for few hundred (average) counts [10]. In the Hahn-echo experiments shown here, we reduced the impact of losses to negligible contribution, by averaging the outcome for each $C_0(\hat{H},t,t')$ across $M = 3000000$ repetitions. Thus, we are left with the Poissonian-distributed counts with a variance of $\sigma_P = C_0$ This noise maps well on a binomial noise model subtending the processing of data $d \in \{0,1\}$. Such a noise model can be naturally implemented in the Bayesian-inference process adopted here [8], in absence of majority voting correction schemes [10].

Quantum projection noise. Quantum projection noise is an unavoidable source of noise when performing projective measurements of a quantum system onto a set of stable states. In this work, the NV centre evolved state with each Hahn-sequences (after tracing out environmental degrees of freedom) is a single-qubit state $|\psi\rangle'_{\text{sys}}$, and the final measurement operation is a projective measurement on the computational basis $\{|0\rangle,|1\rangle\}$, repeated across $M = 3\times10^6$ independent instances, leading to C_0 counts. Under these measurement parameters, ignoring losses already mentioned, the quantum projection noise can be interpreted as a binomial distribution whereby C_0 has a variance $\sigma_b^2(0) = Mp_0(1-p_0)$ [51], where we contracted $p_0 := \mathcal{L}(0|t,t')$. This noise source is then dependent upon the specific p_0 at the end of each evolution, similarly to the Poissonian-noise, but once losses are taken into account, we can observe the upper bound $\sigma_b^2 < \sigma_P^2$. Therefore, the same considerations already done for the optical readout noise apply here as well.

Even if, in principle, the QMLA can be employed to perform noise modelling as well, we observe how it is here impossible to disambiguate all sources of noise. For example quantum projection noise and optical readout are respectively state dependent and readout–counts dependent, but given that here we experimentally access information about the evolved state optically and we have no independent measurement, the two contributions must be considered simultaneously. Similarly, systematic and random errors affecting the probe state preparation, the controlled evolution and the final rotation of the evolved state cannot be effortlessly deconvoluted in the absence of dedicated measurements. Such experiments can well be designed in future investigations adopting QMLA as a characterisation tool, where systematic trusting of parts of the setup leads to the characterisation of others, in the spirit of a modular characterisation [3].

References

1. A. Peruzzo *et al.* A variational eigenvalue solver on a photonic quantum processor. *Nature communications* **5,** 4213 (2014).
2. C. E. Granade, C Ferrie, N Wiebe & D. G. Cory. Robust online Hamiltonian learning. *New Journal of Physics* **14,** 103013 (Oct. 2012).
3. N Wiebe, C Granade, C Ferrie & D. G. Cory. Hamiltonian Learning and Certification Using Quantum Resources. *Physical Review Letters* **112,** 190501–5 (May 2014).
4. N. Wiebe, C. Granade, C. Ferrie & D. Cory. Quantum Hamiltonian learning using imperfect quantum resources. *Physical Review A* **89,** 042314 (2014).
5. J. Wang *et al.* Experimental quantum Hamiltonian learning. *Nature Physics* **13,** 551 (2017).
6. P.-Y. Hou *et al.* Experimental Hamiltonian Learning of an 11-Qubit Solid-State Quantum Spin Register*. *Chinese Physics Letters* **36,** 100303. ISSN: 1741-3540. http://dx.doi.org/10.1088/0256-307x/36/10/100303 (2019).

7. J. Liu & M. West. in *Sequential Monte Carlo algorithms for a class of outer measures* 197–223 (Springer New York, New York, NY, 2001).
8. C. Granade *et al.* QInfer: Statistical inference software for quantum applications. *Quantum* **1,** 5 (2017).
9. H.-P. Breuer & F. Petruccione. *The theory of open quantum systems* (Oxford University Press on Demand, 2002).
10. R. Santagati *et al.* Magnetic-Field Learning Using a Single Electronic Spin in Diamond with One-Photon Readout at Room Temperature. *Phys. Rev. X* **9,** 021019. https://link.aps.org/doi/10.1103/PhysRevX.9.021019 (2 2019).
11. R. E. Kass & A. E. Raftery. Bayes factors. *Journal of the american statistical association* **90,** 773–795 (1995).
12. C. Granade & N. Wiebe. Structured filtering. *New Journal of Physics* **19,** 1–22. ISSN: 13672630 (2017).
13. N. Brunner *et al.* Testing the Dimension of Hilbert Spaces. *Phys. Rev. Lett.* **100,** 210503. https://link.aps.org/doi/10.1103/PhysRevLett.100.210503 (21 2008).
14. G. Carleo & M. Troyer. Solving the quantum many-body problem with artificial neural networks. *Science* **355,** 602–606. ISSN: 0036-8075 (2017).
15. T. Meinhardt, M. Moeller, C. Hazirbas & D. Cremers. *Learning Proximal Operators: Using Denoising Networks for Regularizing Inverse Imaging Problems* in *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 1799–1808.
16. R. Iten, T. Metger, H. Wilming, L. del Rio & R. Renner. Discovering physical concepts with neural networks. *Phys. Rev. Lett. arXiv preprint arXiv:1807.10300* (2018).
17. M. Hutson. Artificial intelligence faces reproducibility crisis. *Science* **359,** 725–726. ISSN: 0036-8075 (2018).
18. H. Bruus & K. Flensberg. *Many-Body Quantum Theory in Condensed Matter Physics An Introduction* ISBN: 9780198566335 (Oxford University Press, 2004).
19. D. W. Berry, G. Ahokas, R. Cleve & B. C. Sanders. Efficient Quantum Algorithms for Simulating Sparse Hamiltonians. *Communications in Mathematical Physics* **270,** 359–371. ISSN: 1432-0916. https://doi.org/10.1007/s00220-006-0150-x (2007).
20. A. M. Childs & R. Kothari. *Simulating Sparse Hamiltonians with Star Decompositions* in *Theory of Quantum Computation, Communication, and Cryptography* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011), 94–103. ISBN: 978-3-642-18073-6.
21. R. Santagati *et al.* Witnessing eigenstates for quantum simulation of Hamiltonian spectra. *Science Advances* **4.** https://advances.sciencemag.org/content/4/1/eaap9646 (2018).
22. S. J. Russell & P. Norvig. *Artificial intelligence: a modern approach* (Pearson Education Limited, 2016).
23. S. Sra, S. Nowozin & S. J. Wright. *Optimization for Machine Learning* (MIT Press, 2012).
24. B. B. *Connectivity and Matching. In: Modern Graph Theory. Graduate Texts in Mathematics* (Springer, 1998).
25. K. P. Murphy. *Machine Learning: A Probabilistic Perspective* ISBN: 0262018020, 9780262018029 (The MIT Press, 2012).
26. C. Bishop. *Pattern Recognition and Machine Learning* ISBN: 9780387310732. https://books.google.co.in/books?id=kTNoQgAACAAJ (Springer, 2006).
27. M. A. Nielsen & I. L. Chuang. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2001).
28. M. Paris & J. Rehacek. *Quantum state estimation* (Springer Science & Business Media, 2004).
29. V. Dunjko & H. J. Briegel. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics* **81,** 074001. https://doi.org/10.1088%2F1361-6633%2Faab406 (2018).
30. G. Torlai *et al.* Neural-network quantum state tomography. *Nature Physics* **14,** 447. https://www.nature.com/articles/s41567-018-0048-5 (2018).
31. M. Kech. Dynamical quantum tomography. *Journal of Mathematical Physics* **57,** 122201 (2016).
32. K. Banaszek, M. Cramer & D. Gross. Focus on quantum tomography. *New Journal of Physics* **15,** 125020 (2013).
33. M. Mohseni, A. Rezakhani & D. Lidar. Quantum-process tomography: Resource analysis of different strategies. *Physical Review A* **77,** 032322 (2008).
34. A. Shabani *et al.* Efficient measurement of quantum dynamics via compressive sensing. *Physical Review Letters* **106,** 1–9. ISSN: 00319007. arXiv: 0910.5498 (2011).
35. K. Rudinger & R. Joynt. Compressed sensing for Hamiltonian reconstruction. *Physical Review A - Atomic, Molecular, and Optical Physics* **92.** ISSN: 10941622. arXiv: 1410.3029 (2015).
36. E. Chertkov & B. K. Clark. Computational inverse method for constructing spaces of quantum models from wave functions. *Physical Review X* **8,** 031029 (2018).
37. A. Shabani, M. Mohseni, S. Lloyd, R. L. Kosut & H. Rabitz. Estimation of many-body quantum Hamiltonians via compressive sensing. *Physical Review A* **84,** 012107 (2011).
38. E. Bairey, I. Arad & N. H. Lindner. Learning a local Hamiltonian from local measurements. *Physical review letters* **122,** 020504 (2019).
39. L. Childress *et al.* Coherent Dynamics of Coupled Electron and Nuclear Spin Qubits in Diamond. *Science* **314,** 281–285 (2006).

40. E. Scerri, E. M. Gauger & C. Bonato. Extending qubit coherence by adaptive quantum environment learning. *New Journal of Physics* **22,** 035002. https://doi.org/10.1088%2F1367-2630%2Fab7bf3 (2020).
41. Z. Zhu, Z. Ji, X. Fan & J. L. Kuo. *Memetic figure selection for cluster expansion in binary alloy systems* in *IEEE SSCI 2011 - Symposium Series on Computational Intelligence - MC 2011: 2011 IEEE Workshop on Memetic Computing* (2011), 15–20. ISBN: 9781612840666. https://ieeexplore.ieee.org/abstract/document/5953635/.
42. A. Gali, M. Fyta & E. Kaxiras. Ab initio supercell calculations on nitrogen-vacancy center in diamond: Electronic structure and hyperfine tensors. *Phys. Rev. B* **77,** 1–12. ISSN: 1098-0121. http://link.aps.org/doi/10.1103/PhysRevB.77.155206 (2008).
43. M. Schlosshauer. Decoherence, the measurement problem, and interpretations of quantum mechanics. *Rev. Mod. Phys.* **76,** 1267–1305. https://link.aps.org/doi/10.1103/RevModPhys.76.1267 (4 2005).
44. U. Weiss. *Quantum Dissipative Systems* 4th. eprint: https://www.worldscientific.com/doi/pdf/10.1142/8334. https://www.worldscientific.com/doi/abs/10.1142/8334 (WORLD SCIENTIFIC, 2012).
45. L. Chirolli & G. Burkard. Decoherence in solid-state qubits. *Advances in Physics* **57,** 225–285. https://doi.org/10.1080/00018730802218067 (2008).
46. X.-L. Deng, D. Porras & J. I. Cirac. Effective spin quantum phases in systems of trapped ions. *Phys. Rev. A* **72,** 063407. https://link.aps.org/doi/10.1103/PhysRevA.72.063407 (6 2005).
47. M. Blok *et al.* Manipulating a qubit through the backaction of sequential partial measurements and real-time feedback. *Nature Physics* **10,** 189 (2014).
48. M. G. Dutt *et al.* Quantum register based on individual electronic and nuclear spin qubits in diamond. *Science* **316,** 1312–1316 (2007).
49. C Ferrie, C. E. Granade & D. G. Cory. How to best sample a periodic probability distribution, or on the accuracy of Hamiltonian finding strategies. *Quantum Information Processing* **12,** 611–623 (Apr. 2012).
50. I. Hincks, C. Granade & D. G. Cory. Statistical inference with quantum measurements: methodologies for nitrogen vacancy centers in diamond. *New Journal of Physics* **20,** 013022 (2018).
51. W. M. Itano *et al.* Quantum projection noise: Population fluctuations in two-level systems. *Phys. Rev. A* **47,** 3554–3570. https://link.aps.org/doi/10.1103/PhysRevA.47.3554 (5 1993).