Reinforcement Learning in Different Phases of Quantum Control

Marin Bukov,[1, *] Alexandre G.R. Day,[1, †] Dries Sels,[1, 2] Phillip Weinberg,[1] Anatoli Polkovnikov,[1] and Pankaj Mehta[1]

[1]*Department of Physics, Boston University, 590 Commonwealth Ave., Boston, MA 02215, USA*
[2]*Theory of quantum and complex systems, Universiteit Antwerpen, B-2610 Antwerpen, Belgium*
(Dated: October 25, 2019)

The ability to prepare a physical system in a desired quantum state is central to many areas of physics such as nuclear magnetic resonance, cold atoms, and quantum computing. Preparing states quickly and with high fidelity remains a formidable challenge. In this work we implement cutting-edge Reinforcement Learning (RL) techniques and show that their performance rivals gradient-based optimal control methods at the task of finding short, high-fidelity driving protocols from an initial to a target state in non-integrable many-body quantum systems of interacting qubits. The quantum state preparation problem, viewed as an optimization problem, is shown to exhibit a spin-glass-like phase transition in the space of protocols as a function of the protocol duration. Our RL-aided approach helps identify variational protocols with nearly optimal fidelity, even in the glassy phase, where optimal state preparation appears exponentially hard. This study highlights the usefulness of RL for applications in out-of-equilibrium quantum physics.

Reliable quantum state manipulation is essential for many areas of physics ranging from nuclear magnetic resonance experiments [1] and cold atomic systems [2, 3] to trapped ions [4–6], quantum optics [7], superconducting qubits [8], nitrogen vacancy centers [9], and quantum computing [10]. However, finding optimal control sequences for such platforms presents a formidable challenge due to our limited theoretical understanding of nonequilibrium quantum systems and the intrinsic complexity of simulating large quantum many-body systems.

For long times, adiabatic state preparation can be used to robustly prepare quantum state provided the change of the Hamiltonian is slow compared to the minimum energy gap. Remarkably, under the assumption of full control and with no constraints on the control fields, one can even prove that solving any quantum optimal control problem is simple [11], i.e. all local optima are also global optima. Unfortunately, both assumptions are often violated in real-life applications. Typical experiments often have stringent constraints on control parameters, such as a maximum magnetic-field strength or a maximal switching frequency. Moreover, decoherence phenomena impose insurmountable time constraints beyond which quantum information is lost irreversibly. For this reason, many experimentally-relevant systems are in practice uncontrollable, i.e. there are no finite-time protocols, which prepare the desired state with unit fidelity. In fact, in Anderson and many-body localized, or periodically-driven systems, which are naturally away from equilibrium, the adiabatic limit does not even exist [12, 13]. This has motivated numerous approaches to quantum state manipulation [14–31]. Despite all advances, at present date surprisingly little is known about how to successfully load a generic interacting quantum system into a desired target state, especially at short times, or even when this is

feasible in the first place [29, 32–34].

In this paper, we adopt a radically different approach to this problem based on machine learning (ML) [36–40]. ML has recently been applied successfully to several problems in equilibrium condensed matter physics [41, 42] and turbulent dynamics [43], and here we demonstrate that Reinforcement Learning (RL) provides deep insights into nonequilibrium quantum dynamics. Specifically, we use a modified version of the Watkins Q-learning algorithm [36] to teach a computer agent to find driving protocols which prepare a quantum system in a target state $|\psi_*\rangle$ starting from an initial state $|\psi_i\rangle$ by controlling a time-dependent field. A far-reaching consequence of our study is the existence of phase transitions in the quantum control landscape of the generic many-body quantum control problem. The glassy nature of the prevalent phase implies that the optimal protocol is exponentially difficult to find. However, as we demonstrate, the optimal solution is unstable to local perturbations. Instead, we discover classes of RL-motivated stable suboptimal protocols[44], the performance of which rival that of the optimal solution in almost every aspect. Analyzing these suboptimal protocols, we construct a variational theory, which demonstrates that the behaviour of exponentially many degrees of freedom (d.o.f.) in a generic many-body quantum system can be described by only a few effective variables. We benchmark the RL results using Stochastic Descent (SD), and compare them to state-of-the-art optimal control methods such as CRAB [29] and GRAPE [35, 45].

In stark contrast to most approaches to quantum optimal control, RL is a *model-free* feedback-control method which allows to find controls even when accurate models of the system are unknown, or parameters in the model are uncertain. A clear advantage of RL over traditional gradient-based optimal control approaches is the fine balance between exploitation of already obtained knowledge and exploration in uncharted parts of the control landscape. Below the quantum speed limit [46], exploration becomes vital and offers an alternative to the prevalent
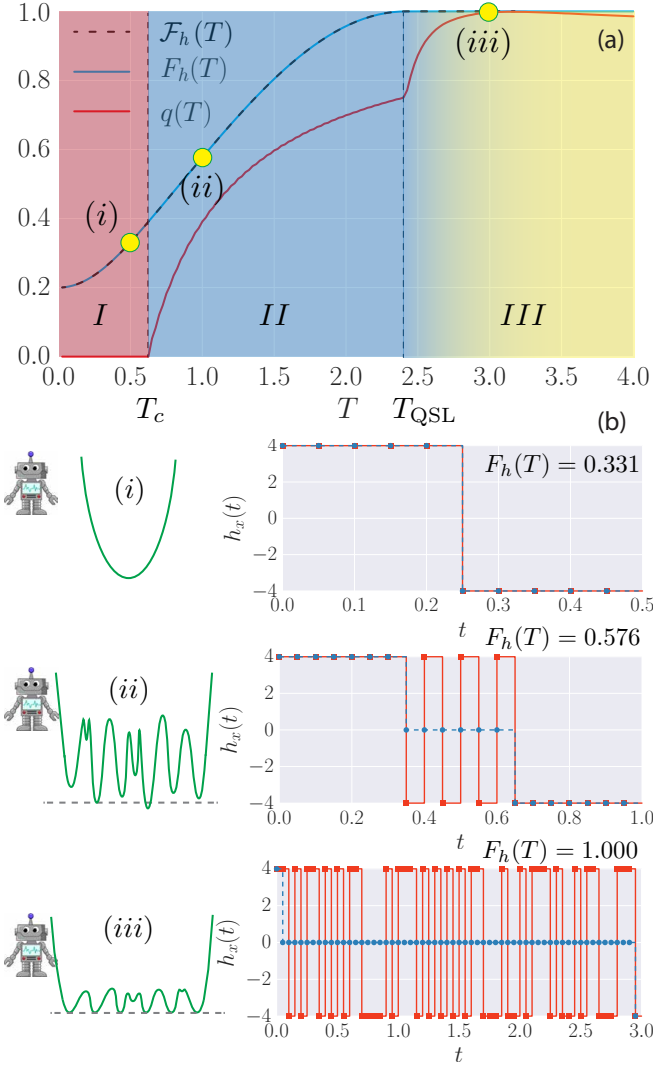
FIG. 1: (a) Phase diagram of the quantum state preparation problem for the qubit in Eq. (1) vs. protocol duration T, as determined by the order parameter $q(T)$ (red) and the maximum possible achievable fidelity $F_h(T)$ (blue), compared to the variational fidelity $\mathcal{F}_h(T)$ (black, dashed). Increasing the total protocol time T, we go from an overconstrained phase I, through a glassy phase II, to a controllable phase III. (b) Left: the infidelity landscape is shown schematically (green). Right: the optimal bang-bang protocol found by the RL agent at the points (i)–(iii) (red) and the variational protocol [35] (blue, dashed).

tage in controlling realistic systems where constructing a reliable effective model is infeasible, for example due to disorder or dislocations.

To manipulate the quantum system, our computer agent constructs piecewise-constant protocols of duration T by choosing a drive protocol strength $h_x(t)$ at each time $t = n\delta t$, $n = \{0, 1, \cdots, T/\delta t\}$, with δt the time-step size. In order to make the agent learn, it is given a reward for every protocol it constructs – the fidelity $F_h(T) = |\langle\psi_*|\psi(T)\rangle|^2$ for being in the target state after time T following the protocol $h_x(t)$ *under unitary Schrödinger evolution*. The goal of the agent is to maximize the reward in a series of attempts. Deprived of any knowledge about the underlying physical model, the agent collects information about already tried protocols, based on which it constructs new, improved protocols through a sophisticated biased sampling algorithm [35]. In realistic applications, one does not have access to infinite control fields; for this reason, we restrict to fields $h_x(t) \in [-4, 4]$, see Fig. 1b. Pontryagin's maximum principle further allows us to focus on bang-bang protocols (Fig. 1b, red), where $h_x(t) \in \{\pm 4\}$, although we verified that RL also works for quasi-continuous protocols with many different steps δh_x [35]. Even though there is only one control field, the space of available protocols grows exponentially with the inverse step size δt^{-1}.

Control Phases of Constrained Qubit Manipulation.— To benchmark the application of RL to physics problems, consider first a two-level system described by

$$H(t) = -S^z - h_x(t)S^x, \qquad (1)$$

where S^α, are the spin-1/2 operators. This Hamiltonian comprises both integrable many-body and non-interacting translational invariant systems, such as the transverse-field Ising model, graphene and topological insulators. The initial $|\psi_i\rangle$ and target $|\psi_*\rangle$ states are chosen as the ground states of (1) at $h_x = -2$ and $h_x = 2$, respectively. Although there exists an analytical solution to solve for the optimal protocol in this case [46], it does not generalize to non-integrable many-body systems. Thus, studying this problem using RL serves a two-fold purpose: (i) we benchmark the protocols obtained by the RL agent demonstrating that, even though RL is a completely model-free algorithm, it still finds the physically meaningful solutions by constructing a minimalistic effective model on-the-fly. The learning process is shown in this Movie; (ii) We reveal an important novel perspective on the complexity of quantum state preparation which, as we show below, generalizes to many-particle systems.

For fixed total ramp time T, the infidelity $h_x(t) \mapsto I_h(T) = 1 - F_h(T)$ represents a "potential landscape", the global minimum of which corresponds to the optimal driving protocol. For bang-bang protocols, the problem of finding the optimal protocol becomes equivalent to finding the ground state configuration of a classical Ising model with complicated interactions. We map out the landscape of local infidelity minima $\{h_x^\alpha(t)\}$ using SD, starting from random bang-bang protocol configu-

paradigm of multi-starting local gradient optimizers [47]. Unlike these methods, the RL agent progressively learns to build a model of the optimization landscape in such a way that the protocols it finds are stable to sampling noise. In this regard, RL-based approaches are particularly well-suited to work with experimental data [48, 49] and, unlike many optimal control methods, they do not require explicit knowledge of local gradients of the control landscape [35, 45]. This offers a considerable advan-

rations [35]. To study the correlations between the infidelity minima as a function of the total ramp time T, we define the correlation $q(T)$, closely related to the Edwards-Anderson order parameter for the existence of spin glass order [50, 51], as

$$q(T) = \frac{1}{16N_T}\overline{\sum_{n=1}^{N_T}\{h_x(n\delta t) - \overline{h_x}(n\delta t)\}^2},\qquad (2)$$

where $\overline{h_x}(t) = N_{\text{real}}^{-1}\sum_{\alpha=1}^{N_{\text{real}}}h_x^\alpha(t)$ is the sample-averaged protocol. If the minima $\{h_x^\alpha(t)\}_{\alpha=1}^{N_{\text{real}}}$ are all uncorrelated, then $\overline{h_x}(t) \equiv 0$, and thus $q(T) = 1$. On the other hand, if the infidelity landscape contains only one minimum, then $\overline{h_x}(t) \equiv h_x(t)$ and $q(T) = 0$. The behaviour of $q(T)$, the maximum fidelity $F_h(T)$ together with a qualitative description of the corresponding infidelity landscapes are shown in Fig. 1.

The control problem for the constrained qubit exhibits three distinct phases as a function of the protocol duration T. If T is greater than the quantum speed limit $T_{\text{QSL}} \approx 2.4$, one can construct infinitely many protocols which prepare the target state with unit fidelity, and the problem is in the *controllable* phase III, c.f. Fig. 1. The red line in Fig. 1b (iii) shows an optimal protocol of unit fidelity found by the agent, whose Bloch sphere representation can be seen in Movie-(iii). In this phase, there is a proliferation of *exactly degenerate*, uncorrelated global infidelity minima, corresponding to protocols of unit fidelity, and the optimization task is easy.

At $T = T_{\text{QSL}}$, the order parameter $q(T)$ exhibits a non-analyticity, and the system undergoes a continuous phase transition to a glassy phase II. For times smaller than T_{QSL} but greater than T_c, the degenerate minima of the infidelity landscape recede to form a glassy landscape with many *non-degenerate* local minima, as reflected by the finite value of the order parameter $0 < q(T) < 1$. As a consequence of this glassy phase, there no longer exists a protocol to prepare the target state with unit fidelity, since it is physically impossible to reach the target state while obeying all constraints. The infidelity minimization problem is non-convex, and determining the best achievable (i.e. optimal) fidelity [a.k.a. the global minimum] becomes a extremely difficult. Figure 1b (ii)shows the best bang-bang protocol found by our computer agent (see Movie-(ii) and [35] for protocols with quasi-continuous actions). This protocol has a remarkable feature: without any prior knowledge about the intermediate quantum state nor its Bloch sphere representation, the model-free RL agent discovers that it is advantageous to first bring the state to the equator – which is a geodesic – and then effectively turns off the control field $h_x(t)$, to enable the fastest possible precession about the z-axis[59]. After staying on the equator for as long as optimal, the agent rotates as fast as it can to bring the state as close as possible to the target, thus optimizing the final fidelity for the available protocol duration.

Decreasing the total ramp time T further, we find a second critical time $T_c \approx 0.6$. For $T < T_c$, $q(T) \equiv 0$ and
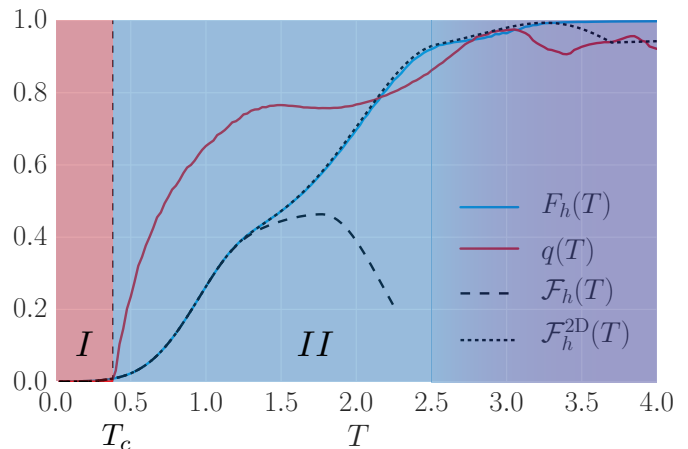

FIG. 2: Phase diagram of the many-body quantum state preparation problem. The order parameter (red) shows a kink at the critical time $T_c \approx 0.4$ when a phase transition occurs from an overconstrained phase (I) to a glassy phase (II). The best fidelity $F_h(T)$ (blue) obtained using SD is compared to the variational fidelity $\mathcal{F}_h(T)$ (dashed) and the 2D-variational fidelity $\mathcal{F}_h^{2D}(T)$ (dotted) [35].

the problem has a unique solution, suggesting that the infidelity landscape is convex. This *overconstrained phase* is labelled I in the phase diagram (Fig. 1a). For $T < T_c$, there exists a unique optimal protocol, even though the achievable fidelity can be quite limited, see Fig. 1b (i) and Movie-(i). Since the state precession speed towards the equator depends on the maximum possible allowed field strength h_x, it follows that $T_c \to 0$ for $|h_x| \to \infty$.

Many Coupled Qubits.—Our results raise the natural question of how much more difficult state preparation is in more complex quantum models. To this end, consider a closed chain of L coupled qubits, which can be experimentally realized with superconducting qubits [8], cold atoms [52] and trapped ions [6]:

$$H(t) = -\sum_{j=1}^{L}\left(S_{j+1}^z S_j^z + h_z S_j^z + h_x(t)S_j^x\right). \qquad (3)$$

We set $h_z = 1$ to avoid the anti-ferromagnet to paramagnet phase transition, and choose the paramagnetic ground states of Eq. (3) at fields $h_x = -2$ and $h_x = 2$ for the initial and target state, respectively. The details of the control field $h_x(t)$ are the same as in the single qubit case, and we use the *many-body* fidelity both the reward and the measure of performance.

Figure 2 shows the phase diagram of the coupled qubits model. First, notice that while the overconstrained-to-glass critical point T_c survives, the quantum speed limit critical point T_{QSL} is (if existent at all) outside the short protocol-time range of interest. Thus, the glassy phase extends over to long and probably infinite ramp times, which offers an alternative explanation for the difficulty of preparing many-body states with high fidelity. Second, observe that, even though unit fidelity is no longer

achievable, there exist nearly optimal protocols with extremely high many-body fidelity[60] at short ramp times. This fact is striking because the Hilbert space of our system grows *exponentially* with L and we are using only one control field to manipulate exponentially many degrees of freedom. Another remarkable characteristic of the optimal solution is that for the system sizes $L \geq 6$ both $q(T)$ and $-1/L \log F_h(T)$ converge to their thermodynamic limit with no visible finite size corrections [35]. This is related to the Lieb-Robinson bound for information propagation which suggests that information should spread over approximately $JT = 4$ sites for the ramp times considered.

Variational Theory.—An additional feature of the optimal bang-bang solution found by the agent is that the entanglement entropy of the half system generated during the evolution always remains small, satisfying an area law [35]. This implies that the system likely follows the ground state of some local, yet a-priori unknown Hamiltonian [53]. This emergent behavior motivated us to use the best protocols found by ML to construct simple variational protocols consisting of a few bangs.

For the single qubit example, these variational protocols are shown in Fig. 1b (dashed blue lines), see also Movies 4-6 for their Bloch sphere representation. The variational fidelity $\mathcal{F}_h(T)$ agrees nearly perfectly with the optimal fidelity $F_h(T)$ obtained using ML techniques, cf. Fig. 1a. We further demonstrate that our variational theory fully captures the physics of the two critical points T_c and T_{QSL} [35]. Interestingly, the variational solution for the single qubit problem coincides with the global minimum of the infidelity landscape all the way up to the quantum speed limit [35].

In the many-body case, the same one-parameter variational ansatz only describes the behaviour in the overconstrained phase, cf. Fig. 2 (dashed line), up to and including the critical point T_c, but fails for $T > T_c$. Nevertheless, a slightly modified, two-parameter variational ansatz, motivated again by the solutions found by the ML agent (see Movie), appears to be fully sufficient to capture the essential features of the optimal protocol much deeper into the glassy phase, as shown by the $\mathcal{F}_h^{\text{2D}}(T)$ curve in Fig. 2. This many-body variational theory features an additional pulse, reminiscent of spin-echo, which appears to control and suppress the generation of entanglement entropy during the drive [35]. Indeed, while the two-parameter ansatz is strictly better than the single-parameter protocol for all $T > T_c$, the difference between the two grows slowly as a function of time. It is only at a later time, $T \approx 1.3$, that the effect of the second pulse really kicks in, and we observe the largest entanglement in the system for the optimal protocol.

Using RL we have thus identified nearly-optimal control protocols [44] which can be parametrized by a few d.o.f. Such simple protocols have been proven to exist in weakly-entangled one-dimensional spin chains [34]. However, the proof of the existence does not imply that these d.o.f. are easy to identify. Initially, the RL agent is com-
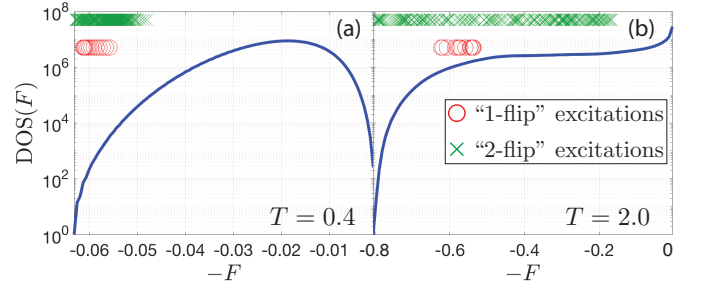

FIG. 3: Density of states (protocols) in the overconstrained phase at $T = 0.4$ (a) and the glassy phase at $T = 2.0$ (b) as a function of the fidelity F. The red circles and the green crosses show the fidelity of the "1-spin" flip and "2-spin" flip excitation protocols above the absolute ground state (i.e. the optimal protocol). The system size is $L = 6$ and each protocol has $N_T = 28$ bangs.

pletely ignorant about the problem and explores many different protocols, while it tries to learn the relevant features. In contrast, optimal control methods, such as CRAB [29], usually have a much more rigid framework, where the d.o.f. of the method are fixed from the beginning. This can limit the performance of those methods below the quantum speed limit [35, 47].

Glassy Behaviour.— It is quite surprising that the dynamics of a non-integrable many-body quantum system, associated with the optimal protocol, is so efficiently captured by such a simple, two-parameter variational protocol, even in the regimes where there is no obvious small parameter and where spin-spin interactions play a significant role. Upon closer comparison of the variational and the optimal fidelities, one can find regions in the glassy phase where the simple variational protocol outperforms the numerical 'best' fidelity, cf. Fig. 2.

To better understand this behavior, we choose a grid of $N_T = 28$ equally-spaced time steps, and compute all 2^{28} bang-bang protocols and their fidelities. The corresponding density of states (DOS) in fidelity space is shown in Fig. 3 for two choices of T in the overconstrained and glassy phase. This allows us to unambiguously determine the ground state of the infidelity landscape (i.e. the optimal protocol). Starting from this ground state, we then construct all excitations generated by local in time flips of the bangs of the optimal protocol. The fidelity of the "1-flip" excitations is shown using red circles in Fig. 3. Notice how, in the glassy phase, these 28 excitations have relatively low fidelities compared to the ground state, and are surrounded by $\sim 10^6$ other states. This has profound consequences: as we are 'cooling' down in the glassy phase, searching for the optimal protocol and coming from a state high up in the infidelity landscape, if we miss one of the 28 elementary excitations, it becomes virtually impossible to reach the global ground state and the situation becomes much worse if we increase the number of steps N_T. On the contrary, in the overconstrained phase, the smaller value of the DOS at the "1-flip" exci-
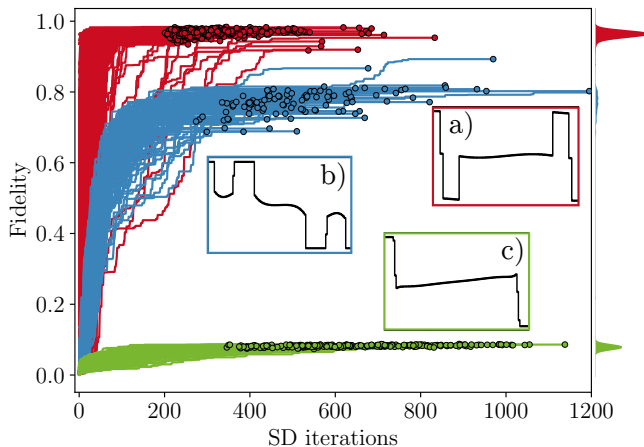
FIG. 4: Fidelity traces of Stochastic Descent (SD) for $T = 3.2, L = 6$ as a function of the number of iterations of the algorithm for 10^3 random initial conditions. The traces are characterized by three main attractors marked by the different colors. The termination of each SD run is indicated by a colored circle. The relative population of the different attractors is shown as a density profile on the right-hand side. Inset (a)-(b)-(c): averaged profile of the protocols obtained for the red, blue and green attractor respectively.

tation ($\sim 10^2$) makes it easier to reach the ground state.

The green crosses in Fig. 3 show the fidelity of the "2-flip" excitations. By the above argument, a "2-flip" algorithm would not see the phase as a glass for $T \lesssim 2.5$, yet it does so for $T \gtrsim 2.5$, marked by the different shading in Fig. 2. Correlated with this observation, we find a signature of a transition also in the improved two-parameter variational theory in the glassy phase [35]. In general, we expect the glassy phase to exhibit a series of phase transitions, reminiscent of the random k-SAT problems [54, 55].

In contrast to the single-qubit system, there are also multiple attractors present in the glassy phase of the many-body system [see Fig. 4]. Each attractor has a typical representative protocol [Fig. 4 inset]. While intra-attractor protocols share the same averaged profile, they can nevertheless have a small mutual overlap – comparable to the overlap of inter-attractor protocols. This indicates that in order to move in between protocols within an attractor, highly non-local moves are necessary. For this reason, GRAPE [45], an algorithm which performs global updates on the protocol by computing exact gradients in the control landscape, also performs very well on our optimisation problem. Similar to SD, in the glassy phase GRAPE cannot escape local minima in the infidelity landscape and, therefore, the same three attractors are found with comparable relative populations to SD, but intra-attractor fluctuations are significantly suppressed due to GRAPE's non-local character.

Outlook.—The appearance of a glassy phase, which

dominates the many-body physics, in the space of protocols of the quantum state preparation problem, has far-reaching consequences for condensed matter physics. Quantum computing relies heavily on our ability to prepare states with high fidelity, yet finding high efficiency state manipulation routines remains a difficult problem. Highly controllable quantum emulators, such as ultracold atoms and ions, depend almost entirely on the feasibility to reach the correct target state, before it can be studied. We demonstrated how, a *model-free* RL agent can provide valuable insights in constructing variational theories which capture almost all relevant features of the dynamics generated by the optimal protocol. Unlike the optimal bang-bang protocol, the simpler variational protocol is robust to small perturbations, while giving comparable fidelities. This implies the existence of nearly optimal protocols, which do not suffer from the exponential complexity of finding the global minimum of the entire optimization landscape. Finally, in contrast with optimal control methods such as SGD, GRAPE, and CRAB that assume an exact model of the physical system, the model-free nature of RL ensures that it can be used to design protocols even when our knowledge of the physical systems we wish to control is incomplete or our system is noisy or disordered.

The existence of phase transitions in quantum control problems may have profound consequences beyond physical systems. We suspect that the glassy behavior observed here maybe a generic feature of many control problems and it will be interesting to see if this is indeed the case. It is our hope that given the close connections between optimal control and RL, the physical interpretation of optimization problems in terms of a glassy phase will help in developing novel efficient algorithms and help spur new ideas in RL and artificial intelligence.

[1] L. M. K. Vandersypen and I. L. Chuang, Rev. Mod. Phys. **76**, 1037 (2005).

[2] S. van Frank, M. Bonneau, J. Schmiedmayer, S. Hild, C. Gross, M. Cheneau, I. Bloch, T. Pichler, A. Negretti, T. Calarco, *et al.*, Scientific reports **6** (2016).

[3] P. B. Wigley, P. J. Everitt, A. van den Hengel, J. Bastian, M. A. Sooriyabandara, G. D. McDonald, K. S. Hardman, C. Quinlivan, P. Manju, C. C. Kuhn, *et al.*, Scientific reports **6** (2016).

[4] R. Islam, E. E. Edwards, K. Kim, S. Korenblit, C. Noh, H. Carmichael, G.-D. Lin, L.-M. Duan, C.-C. Joseph Wang, J. K. Freericks, and C. Monroe, Nature Communications **2**, 377 EP (2011), article.

[5] C. Senko, P. Richerme, J. Smith, A. Lee, I. Cohen, A. Retzker, and C. Monroe, Phys. Rev. X **5**, 021026 (2015).

[6] P. Jurcevic, B. P. Lanyon, P. Hauke, C. Hempel, P. Zoller, R. Blatt, and C. F. Roos, Nature **511**, 202 (2014), letter.

[7] C. Sayrin, I. Dotsenko, X. Zhou, B. Peaudecerf, T. Rybarczyk, S. Gleyzes, P. Rouchon, M. Mirrahimi, H. Amini, M. Brune, J.-M. Raimond, and S. Haroche, Nature **477**, 73 (2011).

[8] R. Barends, A. Shabani, L. Lamata, J. Kelly, A. Mezzacapo, U. Las Heras, R. Babbush, A. Fowler, B. Campbell, Y. Chen, *et al.*, Nature **534**, 222 (2016).

[9] B. B. Zhou, A. Baksic, H. Ribeiro, C. G. Yale, F. J. Heremans, P. C. Jerger, A. Auer, G. Burkard, A. A. Clerk, and D. D. Awschalom, Nat Phys **13**, 330 (2017), letter.

[10] M. A. Nielsen and I. Chuang, *Quantum computation and quantum information* (AAPT, 2002).

[11] H. A. Rabitz, M. M. Hsieh, and C. M. Rosenthal, Science **303**, 1998 (2004).

[12] V. Khemani, R. Nandkishore, and S. L. Sondhi, Nature Physics **11**, 560 (2015).

[13] P. Weinberg, M. Bukov, L. D'Alessio, A. Polkovnikov, S. Vajna, and M. Kolodrubetz, Phyics Reports , 1 (2017).

[14] A. Baksic, H. Ribeiro, and A. A. Clerk, Phys. Rev. Lett. **116**, 230503 (2016).

[15] X. Wang, M. Allegra, K. Jacobs, S. Lloyd, C. Lupo, and M. Mohseni, Phys. Rev. Lett. **114**, 170501 (2015).

[16] R. R. Agundez, C. D. Hill, L. C. L. Hollenberg, S. Rogge, and M. Blaauboer, Phys. Rev. A **95**, 012317 (2017).

[17] S. Bao, S. Kleer, R. Wang, and A. Rahmani, arXiv , 1704.01423 (2017).

[18] G. M. Rotskoff, G. E. Crooks, and E. Vanden-Eijnden, Phys. Rev. E **95**, 012148 (2017).

[19] N. Leung, M. Abdelhafez, J. Koch, and D. I. Schuster, arXiv , arXiv:1612.04929v2 (2016).

[20] Z.-C. Yang, A. Rahmani, A. Shabani, H. Neven, and C. Chamon, Phys. Rev. X **7**, 021027 (2017).

[21] C. Jarzynski, Phys. Rev. A **88**, 040101 (2013).

[22] M. Kolodrubetz, D. Sels, P. Mehta, and A. Polkovnikov, Physics Reports , (2017).

[23] D. Sels and A. Polkovnikov, Proceedings of the National Academy of Sciences **114**, E3909 (2017).

[24] S. J. Glaser, T. Schulte-Herbrüggen, M. Sieveking, O. Schedletzky, N. C. Nielsen, O. W. Sørensen, and C. Griesinger, Science **280**, 421 (1998).

[25] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa, Science **288**, 824 (2000).

[26] N. Khaneja, R. Brockett, and S. J. Glaser, Phys. Rev. A **63**, 032308 (2001).

[27] S. E. Sklarz and D. J. Tannor, Phys. Rev. A **66**, 053619 (2002).

[28] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrggen, and S. J. Glaser, Journal of Magnetic Resonance **172**, 296 (2005).

[29] T. Caneva, T. Calarco, and S. Montangero, Phys. Rev. A **84**, 022326 (2011).

[30] E. Zahedinejad, S. Schirmer, and B. C. Sanders, Phys. Rev. A **90**, 032310 (2014).

[31] E. Zahedinejad, J. Ghosh, and B. C. Sanders, Phys. Rev. Lett. **114**, 200502 (2015).

[32] P. Doria, T. Calarco, and S. Montangero, Phys. Rev. Lett. **106**, 190501 (2011).

[33] T. Caneva, A. Silva, R. Fazio, S. Lloyd, T. Calarco, and S. Montangero, Phys. Rev. A **89**, 042322 (2014).

[34] S. Lloyd and S. Montangero, Phys. Rev. Lett. **113**, 010502 (2014).

[35] See Supplemental Material.

[36] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2017).

[37] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn* (2007).

[38] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, Nature **518**, 529 (2015), letter.

[39] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Nature **529**, 484 (2016), article.

[40] V. Dunjko, J. M. Taylor, and H. J. Briegel, Phys. Rev. Lett. **117**, 130501 (2016).

[41] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[42] J. Carrasquilla and R. G. Melko, Nat Phys **13**, 431 (2017).

[43] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, Proceedings of the National Academy of Sciences , 201606075 (2016).

[44] T. R. Gingrich, G. M. Rotskoff, G. E. Crooks, and P. L. Geissler, Proceedings of the National Academy of Sciences , 201606273 (2016).

[45] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrggen, and S. J. Glaser, Journal of Magnetic Resonance **172**, 296 (2005).

[46] G. C. Hegerfeldt, Phys. Rev. Lett. **111**, 260501 (2013).

[47] J. J. W. H. Sørensen, M. K. Pedersen, M. Munch, P. Haikka, J. H. Jensen, T. Planke, M. G. Andreasen, M. Gajdacz, K. Mølmer, A. Lieberoth, and J. F. Sherson, Nature **532**, 210 (2016).

[48] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger, Phys. Rev. Lett. **116**, 090405 (2016).

[49] A. A. Melnikov, H. P. Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, and H. J. Briegel, arXiv , arXiv:1706.00868 (2017).

[50] T. Castellani and A. Cavagna, Journal of Statistical Me-

chanics: Theory and Experiment **2005**, P05012 (2005).

[51] L. O. Hedges, R. L. Jack, J. P. Garrahan, and D. Chandler, Science **323**, 1309 (2009).

[52] J. Simon, W. S. Bakr, R. Ma, M. E. Tai, P. M. Preiss, and M. Greiner, Nature **472**, 307 (2011).

[53] L. Vidmar, D. Iyer, and M. Rigol, arXiv , arXiv:1512.05373v3 (2015).

[54] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).

[55] S. C. Morampudi, B. Hsu, S. L. Sondhi, R. Moessner, and C. R. Laumann, arXiv , arXiv:1704.00238 (2017).

[56] P. Weinberg and M. Bukov, SciPost Phys. **2**, 003 (2017).

[57] M. Tomka, T. Souza, S. Rosenberg, and A. Polkovnikov, arXiv , arXiv:1606.05890 (2016).

[58] C. Jarzynski, S. Deffner, A. Patra, and Y. Subasi, Phys. Rev. E **95**, 032122 (2017).

[59] Notice that the agent does not have control over the h_z field.

[60] We expect the *many-body* overlap between two states to be exponentially small in the system size L.

I. SUPPLEMENTAL MATERIAL

II. GLASSY BEHAVIOUR OF DIFFERENT MACHINE LEARNING AND OPTIMAL CONTROL ALGORITHMS WITH LOCAL AND NONLOCAL FLIP UPDATES

A. Reinforcement Learning

Reinforcement Learning (RL) is a subfield of Machine Learning (ML) where a computer agent is taught to perform and master a specific task by performing a series of actions in order to maximize a reward function (in our case, the final fidelity). As already mentioned in the main text, we use a modified version of Watkins online Q-learning algorithm with linear function approximation and eligibility traces [36] to teach a RL agent to find protocols of optimal fidelity. Below, we briefly summarize the details of the procedure. For a detailed description of the algorithm, we refer the reader to Ref. [36]. A Python implementation of the algorithm is available on Github.

The fidelity optimization problem is defined as an episodic, undiscounted Reinforcement Learning task. Each episode takes a fixed number of steps $N_T = T/\delta t$, where T is the total ramp time, and δt is the physical time step (protocol time step). We define the state $\mathcal{S}$, action $\mathcal{A}$ and reward $\mathcal{R}$ spaces, respectively, as

$$\mathcal{S} = \{s = (t, h_x(t))\}, \qquad \mathcal{A} = \{a = \delta h_x\}, \qquad \mathcal{R} = \{r \in [0,1]\}. \tag{4}$$

The state space consists of all tuples $(t, h_x(t))$ of time t with the corresponding magnetic field $h_x(t)$. Notice that no information about the quantum state whatsoever is encoded in the RL state, and hence the RL algorithm is model-free. Thus, the RL agent is able to learn circumventing the difficulties associated with the notions of quantum physics. Including time t to the state is not a common procedure in Q-learning, but is required in order for the agent to be able to estimate how far away it is from the episode's end.

The action space $\mathcal{A}$ consists of all jumps in the protocol $h_x(t)$, denoted by δh_x. Thus, protocols are constructed as piecewise-constant functions. We restrict the available actions of the RL agent such that at all times the field $h_x(t)$ is in the interval $[-4, 4]$. The bang-bang and quasi-continuous protocols discussed in the next section are examples of the family of protocol functions we allow in the simulation. Pontryagin's maximum principle in optimal control applied to our problem postulates that, given any protocol of certain fidelity, there exists a bang-bang protocol of the same fidelity. Thus, allowing for bang-bang protocols, our study is entitled to capture the optimal solution as well. Generally, this follows from the Trotter-Suzuki decomposition, which implies that any time evolution, not only optimal, can be approximated by a series of bangs with an arbitrary accuracy.

Last but not least, the reward space $\mathcal{R}$ is the space of all real numbers in the interval $[0,1]$. The rewards for the agent are given only at the end of each episode, according to:

$$r(t) = \begin{cases} 0, & \text{if } t < T \\ |\langle \psi_* | \psi(T) \rangle|^2, & \text{if } t = T \end{cases} \tag{5}$$

This reflects the fact that we are not interested in which quantum state the physical system is in during the ramp; all that matters is that we maximize the final fidelity.

An essential part of the RL problem is to define the environment, with which the agent interacts in order to learn. We choose this to consist of the Schrödinger initial value problem, together with the target state:

$$\text{Environment} = \{i\partial_t |\psi(t)\rangle = H(t)|\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_i\rangle, \quad |\psi_*\rangle\}, \tag{6}$$

where $H(t)$ is the Hamiltonian whose time dependence is defined through the magnetic filed $h_x(t)$ which the agent is constructing during the episode via online Q-learning updates.

Let us now briefly illustrate the protocol construction algorithm: for instance, if we start in the initial RL state $s_0 = (t = 0, h = -4)$, and take the action $a = \delta h_x = 8$, we go to the next RL state $s_1 = (\delta t, +4)$. As a result of the interaction with the environment, the initial quantum state is evolved forward in time for one time step (from time $t_0 = 0$ to time $t_1 = \delta t$) with the constant Hamiltonian $H(h = 4)$: $|\psi(\delta t)\rangle = \mathrm{e}^{-iH(h=4)\delta t}|\psi_i\rangle$. After each step we compute the local reward according to Eq. (5), and update the Q-function, even though the instantaneous reward at that step might be zero. This procedure is repeated until the end of the episode is reached at $t = T$. In general, one can imagine this Markov process as a state-action-reward chain

$$s_0 \to a_0 \to r_0 \longrightarrow s_1 \to a_1 \to r_1 \longrightarrow s_2 \to \cdots \longrightarrow s_{N_T}.$$

The total return for a state-action reward chain is defined as the sum of all local rewards during the episode: $R = \sum_{i=0}^{N_T} r_i$. The central object in Watkins Q-Learning is the $Q(s, a)$ function which is given by the expected return at
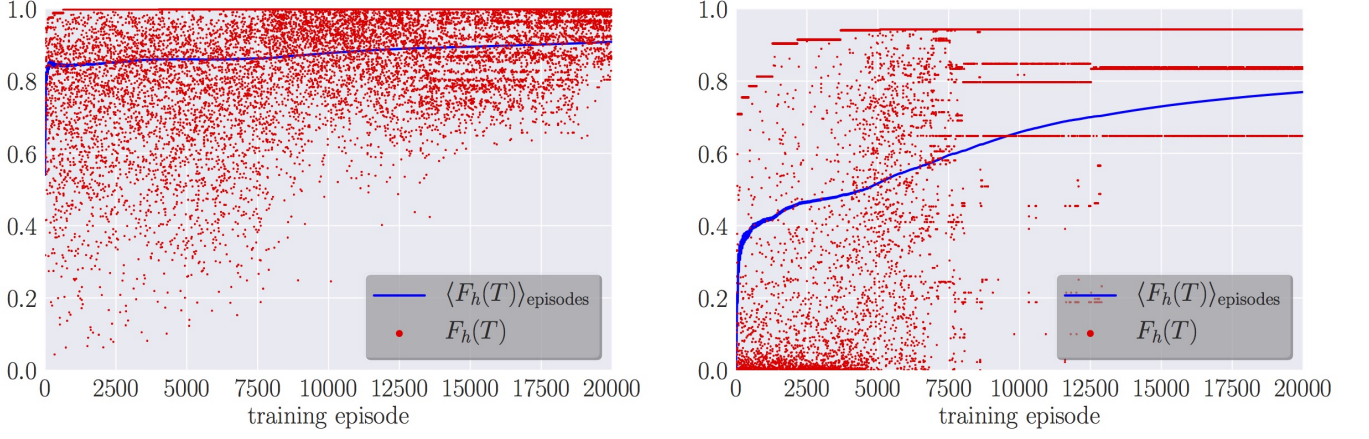
FIG. 5: Learning curves of the RL agent for $L = 1$ at $T = 2.4$ (left) [see Movie] and $L = 10$ at $T = 3.0$ (right) [see Movie]. The red dots show the instantaneous reward (i.e. fidelity) at every episode, while the blue line the cumulative episode-average. The ramp-up of the RL temperature β_{RL} suppresses exploration over time which leads to a gradually increasing average fidelity. The time step is $\delta t = 0.05$.

the end of each episode for starting in a state s and taking the action a. It satisfies the Bellman optimality equation, the solution of which cannot be obtained in a closed form. The usual way of solving the Bellman equation numerically is to make use of temporal difference learning, which results in the following Q-Learning update rule [36]

$$Q(s_i, a_i) \longleftarrow Q(s_i, a_i) + \alpha \left[r_i + \max_a Q(s_{i+1}, a) - Q(s_i, a_i) \right], \tag{7}$$

where the learning rate $\alpha \in (0, 1)$. Whenever $\alpha \approx 1$, the convergence of the update rule (7) can be slowed down or even precluded, in cases where the error $\delta_t = r_i + \max_a Q(s_{i+1}, a) - Q(s_i, a_i)$ becomes significant. On the contrary, $\alpha \approx 0$ corresponds to very slow learning. Thus, the optimal value for the learning rate lies in between, and is determined empirically for the problem under consideration.

To allow for the efficient implementation of nearly continuous drives (used to generate the quasi-continuous protocols), we employ a linear function approximation to the Q-function, using equally-spaced tilings along the entire range of $h_x(t) \in [-4, 4]$ [36]. This allows the RL agent to generalize, i.e. gain information about the fidelity of some protocols not yet encountered.

We then iterate the algorithm for 2×10^4 episodes. The exploration-exploitation dilemma [36] requires a fair amount of exploration, in order to ensure that the agent visits large parts of the RL state space which prevents it from getting stuck in a local minimum from the beginning. Too much exploration, and the agent will not be able to learn. On the other hand, no exploration whatsoever guarantees that the agent will repeat deterministically a given strategy, though it will be unclear whether there exists a better, yet unseen one. In the longer run, we cannot preclude the agent from ending up in a local minimum of reward space. In such cases, we run the algorithm multiple times starting from a random initial condition, and post-select the outcome.

Due to extremely large state space, we employed a novel form of replay to ensure that our RL algorithm could learn from the high fidelity paths it encountered. Our replay algorithm alternates between two different ways of training our RL agent which we call training phases: an "exploratory" training-phase where the RL agent exploits the current Q-function to explore, and a "replay" training-phase where we replay the best encountered protocol. This novel form of replay, to the best of our knowledge, has not been used previously. In the exploratory training-phase, which lasts 40 episodes, the agent takes actions according to a softmax probability distribution based on the instantaneous values of the Q-function. In other words, at each time step, the RL agent looks up the instantaneous values $Q(s, :)$ corresponding to all available actions, and computes a probability for each action: $P(a) \sim \exp(\beta_{\mathrm{RL}} Q(s, a))$. The amount of exploration is set by β_{RL}, with $\beta_{\mathrm{RL}} = 0$ corresponding to random actions and $\beta_{\mathrm{RL}} = \infty$ corresponding to always taking greedy actions with respect to the current Q-function. Here we use an external 'learning' temperature scale, the inverse of which, β_{RL}, is linearly ramped down as the number of episodes progresses. In the replay training-phase, which is also 40 episodes, we replay the best-encountered protocol up to the given episode. Through this procedure, when the next exploratory training-phase begins again, the agent is biased to do variations on top of the best-encountered protocol, effectively improving it, until it reaches a reasonably good fidelity. Two learning curves of the RL agent are shown in Fig. 5.

B. Stochastic Descent

To benchmark the results obtained with RL, we use a greedy stochastic descent (SD) algorithm to sample the infidelity landscape minima containing the driving protocols. We restrict our SD algorithm to exploring bang-bang protocols, for which $h_x(t) \in \{\pm 4\}$. The algorithm starts from a random protocol configuration and proposes local field updates at a time t chosen uniformly in the interval $[0, T]$. The updates consist in changing the applied field $h_x(t) \to h'_x(t)$ only if this increases the fidelity. Ideally, the protocol is updated until all possible local field updates can only decrease the fidelity. Practically, for some ramp times, ensuring that a true local minima with respect to 1-flip is reached can be computational expensive. Therefore, we restrict the number of fidelity evaluations to be at most $20 \times T/\delta t$. In this regard, the obtained protocol is a local minimum with respect to local (1-flip) field updates. The stochastic descent is repeated multiple times with different initial random protocols. The set of protocols $\{h^\alpha | \alpha = 1, \ldots, N_{\text{real}}\}$ obtained with stochastic descent is used to calculate the glass-like order parameter $q(T)$ (see main text). A Python implementation of the algorithm is available on Github.

C. CRAB

Chopped RAndom Basis (CRAB) is a state-of-the-art optimal control algorithm designed to tackle many-body quantum systems [29, 32]. The idea behind CRAB is to decompose the unknown driving protocol into a complete basis (Fourier, Laguerre, etc.), and impose a cut-off on the number of 'harmonics' kept for the optimisation. The algorithm then uses an optimiser to find the values for the expansion coefficients, which optimise the cost function of the problem.

Following Ref. [29], we make a Fourier-basis ansatz for the driving protocol.

$$h_{\text{CRAB}}(t) \;=\; h_0(t) \left(1 + \frac{1}{\lambda(t)} \sum_{i=1}^{N_c} A_i \cos \omega_i t + B_i \sin \omega_i t \right), \tag{8}$$

where the Fourier coefficients $\{A_i, B_i, \omega_i\}$ which parametrise the protocol are found using an implementation of the Nelder-Mead optimization method in the SciPy python library. The number of harmonics kept in the optimisation is given by N_c. The CRAB algorithm uses two auxiliary functions, defined by the user: the first function $h_0(t)$ is a trial initial guess ansatz for the protocol, while the second function, $\lambda(t)$ imposes the boundary conditions $\lambda \to \infty$ for $t \to 0$ and $t \to T$ to the Fourier expansion term.

The cost function which we optimise in the state preparation problem

$$\mathcal{C}[h(t)] = \mathcal{F}(\{A_i, B_i, \omega_i\}) + \frac{1}{16T} \int_0^T \mathrm{d}t [h(t)]^2 \tag{9}$$

contains the fidelity $\mathcal{F}(\{A_i, B_i, \omega_i\})$ at the end of the protocol, and an additional penalty coming from the L^2 norm of the protocol to keep the optimal protocols bounded. The last constraint is required for a better and honest comparison with the RL, SD, and GRAPE algorithms.

Applying CRAB to the state preparation problem from the main text systematically, we choose $\lambda(t) = 1/\sin(\pi t/T)^2$, and $h_0(t) = -2 + 4t/T$. We also consider $N_c = 10, 20$ to study how much an effect increasing the number of degrees of freedom will have on the optimal protocols found. For each value of N_c we start the optimization algorithm with 10 random initial configurations. We define the optimal protocol as the protocol with the best fidelity out of that group of 10. The random initial frequencies ω_i are chosen the same way as outlined in Ref. [29] while the amplitudes A_i and B_i are chosen uniformly between -10 and 10.

D. GRAPE

GRadient Ascend Pulse Engineering, is a numeric gradient-based optimal control method, first introduced in the context of NMR spectroscopy [45]. As suggested by its name, the method performs gradient optimization. Instead of restricting the protocols to bang-bang type, the method works with quasi-continuous protocols. Protocol magnitudes can take on any value within the allowed manifold but, unlike CRAB, are piecewise constant in time.

In the present case, one can efficiently compute the gradient of the fidelity as follows. Consider the fidelity for some trial protocol $h_x(t)$,

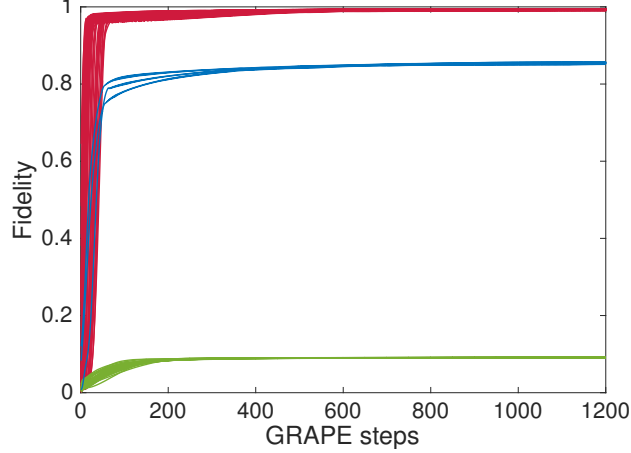$$F_h(T) = |\langle \psi_* | U(T,0) | \psi_i \rangle|^2, \tag{10}$$

FIG. 6: Fidelity traces of GRAPE for $T = 3.2$ and $L = 6$ as a function of the number of gradient ascend steps for 10^2 random initial conditions. This figure should be compared to Fig. 4 in the main text. GRAPE clearly gets attracted by the same three attractors as SD but has much smaller intra-attractor fluctuations, presumably due to the non-locality of the updates and the continuous values of the control field $h_x \in [-4, 4]$ used in GRAPE. This shows that the glass control phase is present for both local and nonlocal update algorithms.

where $U(T, 0)$ denotes the time evolution operator from 0 to T. Let us further decompose the Hamiltonian as $H = H_0 + h_x(t)X$, where H_0 is the part over which we have no control, and X denotes the operator we control. The functional derivative of the fidelity with respect to the protocol thus becomes

$$\frac{\delta F_h(T)}{\delta h_x(t)} = i \langle\psi_*| U(T, 0) |\psi_i\rangle \langle\psi_i| U(0, t)XU(t, T) |\psi_*\rangle - i \langle\psi_*| U(T, t)XU(t, 0) |\psi_i\rangle \langle\psi_i| U(0, T) |\psi_*\rangle \qquad (11)$$

Although this expression appears hard to evaluate, it takes on a very simple form

$$\frac{\delta F_h(T)}{\delta h_x(t)} = 2\text{Im} \left[\langle\phi(t)| X |\psi_i(t)\rangle \right], \qquad (12)$$

where $|\psi_i(t)\rangle = U(t, 0) |\psi_i\rangle$ denotes the initial state propagated to time t and $\langle\phi(t)| = \langle\psi_i(T)|\psi_*\rangle \langle\psi_*| U(T, t)$ denotes the (scaled) final state propagated back in time to time t. Notice that this procedure requires us to exactly know this time evolution operator (i.e. a model of physical system to be controlled). Hence, by propagating both the initial state forward and the target state backward in time one gets access to the full gradient of the control landscape.

To find a local maximum one can simply now gradient ascend the fidelity. A basic algorithm thus goes as follows:

(i) Pick a random initial magnetic field $h_x^0(t)$.

(ii) Compute first $|\psi_i(t)\rangle$ and then $\langle\phi(t)|$ for the current setting of the magnetic field $h_x^N(t)$.

(iii) Update the control field $h_x^{N+1}(t) = h_x^N(t) + \epsilon_N \text{Im} \left[\langle\phi(t)| X |\psi_i(t)\rangle \right]$.

(iv) Repeat (ii) and (iii) until the desired tolerance is reached.

Here ϵ_N is the step size in each iteration. Choosing a proper step size can be a difficult task. In principle the fidelity should go up after each iteration but if the step size it too large, the algorithm can overshoot the maximum, resulting in a worse fidelity. To avoid this, one should adapt the step size during the algorithm. Numerically we have observed that, in order to avoid overshooting saddles/maxima, $\epsilon_N \propto 1/\sqrt{N}$.

E. Comparison between the RL, SD, CRAB and GRAPE

While a detailed comparison between the ML algorithms and other optimal control algorithms is an interesting topic, it is beyond the scope of the present paper. Below, we only show that both the RL algorithm is capable of finding optimal bang-bang protocols for the quantum control problem from the main text, and that its performance rivals that of SD and the state-of-the-art algorithms for many-body quantum problems CRAB and GRAPE.
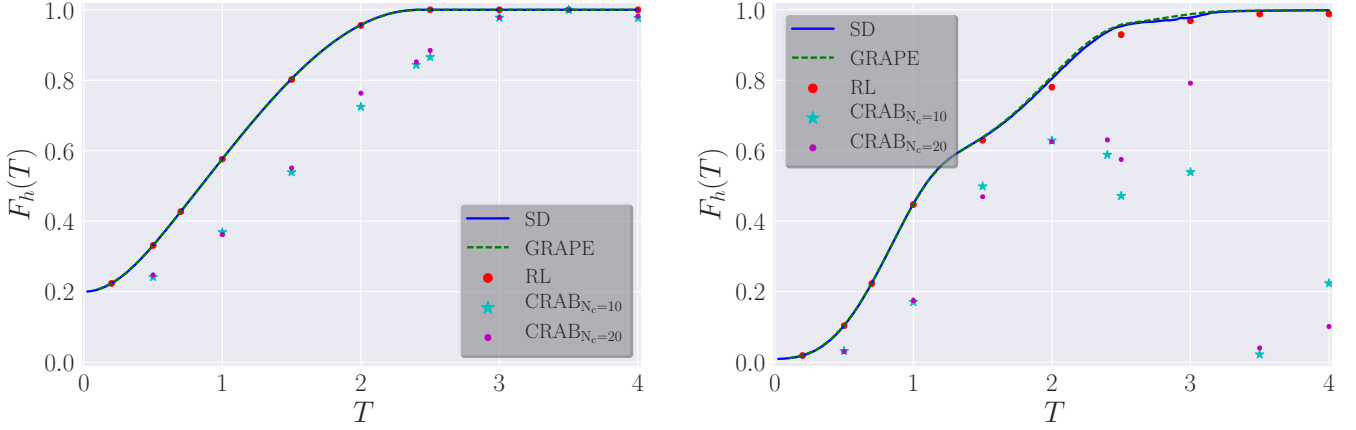
FIG. 7: Comparison between the best fidelities obtained using SD (solid blue line), RL (big red dots), GRAPE (dashed green line) and CRAB (cyan star and small magenta dot) for $L = 1$ (left) and $L = 6$ (right). Here N_c denotes the cap in the number of harmonics kept in the CRAB simulation.

The result for the single qubit $L = 1$ are shown in Fig. 7 (left). The most important points can be summarised as follows:

- RL, GS and GRAPE all find the optimal protocol.

- Below the quantum speed limit, T_{QSL}, CRAB finds good, but clearly suboptimal protocols. The plots also show the glassiness represents a generic feature of the constrained optimization problem and not the method used to perform the optimization. Increasing the cutoff N_c, and with it the number of effective degrees of freedom, does not lead to a sizeable improvement in CRAB. One explanation is the following: for the single qubit, we know that the variational protocol, which contains at most two bangs (see Fig. 1b), is a global minimum of the optimisation landscape. Such a protocol can easily be approximated using up to $N_c = 20$ harmonics. However, allowing more degrees of freedom comes at a huge cost due to the glassiness of the problem: there exist many quasi-degenerate local minima for the algorithm to get stuck in.

The comparison for the Many Coupled Qubits system for $L = 6$ are shown in Fig. 7 (right). Larger values of L do not introduce any change in the behaviour, as we argue in a subsequent section below. In the many-body case, the variational protocol, which contains four bangs, is shown *not* to be the global minimum of the infidelity landscape. Instead, the true global minimum contains many more bangs which, however, only marginally improve the fidelity.

- All algorithms give reasonable fidelities, see Fig. 7 (right)

- Even though GRAPE seems to display better performance out of the four methods, one should not forget that this algorithm, which uses global flips, requires knowledge of all fidelity gradients – valuable information which is not easily accessible through experimental measurements. One has to keep in mind though, that this comparison is not completely honest, since GRAPE allows for the control field h_x to take any value in the interval $[-4, 4]$, which offers a further advantage over the bang-bang based RL. On the other hand, the model-free RL rivals the performance of GRAPE at all ramp times, and outperforms CRABS, even for the single qubit below the quantum speed limit, where the problem enters the glassy phase. The reason for the seemingly slowly decreasing performance of RL with T is that, since the total ramp time is held fixed, the number of bangs increases exponentially with $T \sim N_T$, and hence the state space which has to be explored by the agent also grows exponentially. A detailed study of the scaling is postponed to future studies.

- All algorithms suffer from the glassiness in the optimisation landscape. This is not surprising, since the glass phase is an intrinsic property of the infidelity landscape, as defined by the optimisation problem, and does not depend on which algorithm is used to look for the optimal solution. Fig. 6 shows the fidelity traces as a function of running time for GRAPE. Comparing this to the corresponding results for SD, see Fig. 4, we see a strikingly similar behaviour, even though GRAPE uses nonlocal flip updates in contrast to SD. This means that, in the glassy phase, GRAPE also gets stuck in suboptimal attractors, similar to SD and RL. Thus, as an important consequence, the glassy phase affects both local and nonlocal-update algorithms.
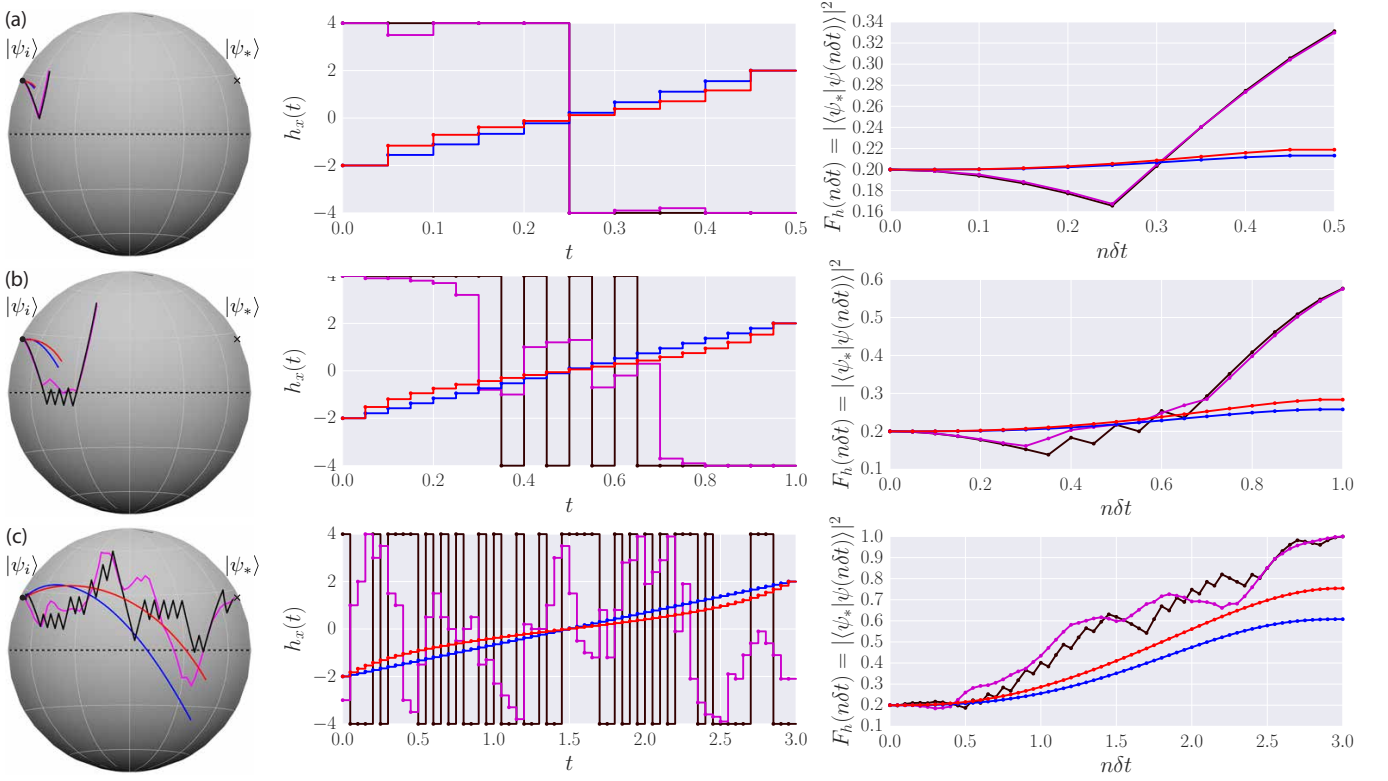
FIG. 8: Comparison between the bang-bang (black) and quasi-continuous (magenta) protocols found by the RL agent, and the Landau-Zener (blue) and geodesic (red) protocols computed from analytical theory in the overconstrained phase for $T = 0.5$ (a), the glassy phase for $T = 1.0$ (b), and the controllable phase for $T = 3.0$ (c). The left column shows the representation of the corresponding protocol on the Bloch sphere, the middle one – the protocols themselves, and the right column – the instantaneous fidelity in the target state $\psi_*\rangle$.

III. PERFORMANCE OF THE DIFFERENT DRIVING PROTOCOLS FOR THE QUBIT

It is interesting to compare the bang-bang and quasi-continuous driving protocols found by the agent to a simple linear protocol, which we refer to as Landau-Zener (LZ), and the geodesic protocol, which optimizes local fidelity close to the adiabatic limit essentially slowing down near the minimum gap [57]. We find that the RL agent offers significantly better solutions in the overconstrained and glassy phases, where the optimal fidelity is always smaller than unity. The Hamiltonian of the qubit together with the initial and target states read:

$$H(t) = -S^z - h_x(t)S^x, \qquad |\psi_i\rangle \sim (-1/2 - \sqrt{5}/2, 1)^\mathsf{T}, \qquad |\psi_*\rangle \sim (1/2 + \sqrt{5}/2, 1)^\mathsf{T}, \qquad (13)$$

where $|\psi_i\rangle$ and $|\psi_*\rangle$ are the ground state of $H(t)$ for $h_i = -2$ and $h_* = +2$ respectively. Note that for bang-bang protocols, the initial and target states are not eigenstates of the control Hamiltonian since $h_x(t)$ takes on the values ± 4.

The RL agent is initiated at the field $h(t = 0) = h_{\min} = -4.0$. The RL protocols are constructed from the following set of jumps, δh_x, allowed at each protocol time step δt:

- *bang-bang* protocol: $\delta h_x \in \{0.0, \pm 8.0\}$ which, together with the initial condition, constrains the field to take the values $h_x(t) \in \{\pm 4.0\}$.

- *quasi-continuous* protocol: $\delta h_x \in \{0.0, \pm 0.1, \pm 0.2, \pm 0.5, \pm 1.0, \pm 2.0, \pm 4.0, \pm 8.0\}$. We restrict the actions available in a state to ensure $h_x(t) \in [-4.0, 4.0]$.

Interestingly, the RL agent figures out that it is always advantageous to first jump to $h_{\max} = +4.0$ before starting the evolution, as a consequence of the positive value of the coefficient in front of S^z.

The analytical adiabatic protocols are required to start and end in the initial and target states, which coincide with the ground states of the Hamiltonians with fields $h_i = -2.0$ and $h_* = 2.0$, respectively. They are defined as follows:

- *Landau-Zener(LZ)* protocol: $h_x(t) = (h_* - h_i)t/T + h_i$

- *geodesic* protocol: $h_x(t) = \tan(at + b)$, where $b = \arctan(h_i)$ and $a = \arctan(h_* - b)/T$.

Figure 8 shows a comparison between these four protocol types for different values of T, corresponding to the three quantum control phases. Due to the instantaneous gap remaining small compared to the total ramp time, the LZ and geodesic protocols are very similar, irrespective of T. The two protocols significantly differ only at large T, where the geodesic protocol significantly outperforms the linear one. An interesting general feature for the short ramp times considered is that the fidelities obtained by the LZ and geodesic protocols are clearly worse than the ones found by the RL agent. This points out the far-from-optimal character of these two approaches, which essentially reward staying close to the instantaneous ground state during time evolution. Looking at the fidelity curves in Fig. 8, we note that, before reaching the optimal fidelity at the end of the ramp for the overconstrained and glassy phases, the instantaneous fidelity drops below its initial value at intermediate times. This suggests that the angle between the initial and target states on the Bloch sphere becomes larger in the process of evolution, before it can be reduced again. Such situation is very reminiscent of counter-diabatic or fast forward driving protocols, where the system can significantly deviate from the instantaneous ground state at intermediate times [9, 23, 58]. Such problems, where the RL agent learns to sacrifice local rewards in view of obtaining a better total reward in the end are of particular interest in RL [36].

IV. CRITICAL SCALING ANALYSIS OF THE CONTROL PHASE TRANSITIONS

In this section, we show convincing evidence for the existence of the phase transitions in the quantum state preparation problem discussed in the main text. We already argued that there is a phase transition in the optimization problem as a function of the protocol time T. Mathematically the problem is formulated as

$$h_x^{\text{optimal}}(t) = \underset{h_x(t):|h_x(t)|\leq 4}{\arg\min} \{I_h(T)\} = \underset{h_x(t):|h_x(t)|\leq 4}{\arg\max} |\langle\psi_*|\mathcal{T}_t e^{-i\int_0^T \mathrm{d}t H[h_x(t)]}|\psi_i\rangle|^2, \tag{14}$$

i.e. as finding the optimal driving protocol $h_x^{\text{optimal}}(t)$, which transforms the initial state (the ground state of the Hamiltonian H at $h_x = -2$) into the target state (the ground state of H corresponding to $h_x = 2$) maximizing the fidelity in ramp time T under unitary Schrödinger evolution. We assume that the ramp protocol is bounded, $h_x(t) \in [-4, 4]$, for all times during the ramp. In this section, we restrict the analysis to bang-bang protocols only, for which $h_x(t) \in \{\pm 4\}$. The minimum protocol time step is denoted by δt. There are two different scaling limits in the problem. We define a continuum limit for the problem as $\delta t \to 0$ while keeping the total ramp time $T = const$. Additionally, there is the conventional thermodynamic limit, where we send the system size $L \to \infty$.

As we already alluded to in the main text, one can think of this optimization problem as a minimization in the infidelity landscape, determined by the mapping $h_x(t) \mapsto I_h(T) = 1 - F_h(T)$, where each protocol is assigned a point in fidelity space – the probability of being in the target state after evolution for a fixed ramp time T. Finding the global minimum of the landscape then corresponds to obtaining the optimal driving protocol for any fixed T.

To obtain the set of local minima $\{h_x^\alpha(t)|\alpha = 1, \ldots, N_{\text{real}}\}$ of the infidelity landscape at a fixed total ramp time T and protocol step size δt, we apply Stochastic Descent(SD), see above, starting from a random protocol configuration, and introduce random *local* changes to the bang-bang protocol shape until the fidelity can no longer be improved. This method is guaranteed to find a set of representative local infidelity minima with respect to "1-flip" dynamics, mapping out the bottom of the landscape of $I_h(T)$. Keeping track of the mean number of fidelity evaluations N_{fidevals} required for this procedure, we obtain a measure for the average time it takes the SD algorithm to settle in a local minimum. While the order parameter $q(T)$ (see below) was used in the main text as a measure for the static properties of the infidelity landscape, dynamic features are revealed by studying the number of fidelity evaluations N_{fidevals}.

As discussed in the main text, the rich phase diagram of the problem can also be studied by looking at the order parameter function q (closely related to the Edwards-Anderson order parameter for detecting glassy order in spin systems [50]):

$$q(T) = \frac{1}{16N_T} \overline{\sum_{n=1}^{N_T} \{h_x(n\delta t) - \overline{h_x}(n\delta t)\}^2}, \qquad \overline{h_x}(t) = \frac{1}{N_{\text{real}}} \sum_{\alpha=1}^{N_{\text{real}}} h_x^\alpha(t). \tag{15}$$

Here, N_T is the total number of protocol time steps of fixed width δt, N_{real} is the total number of random protocol realisations $h_x^\alpha(t)$ probing the minima of the infidelity landscape (see previous paragraph), and the factor $1/16$ serves to normalise the squared bang-bang drive protocol $h_x^2(t)$ within the range $[-1, 1]$.
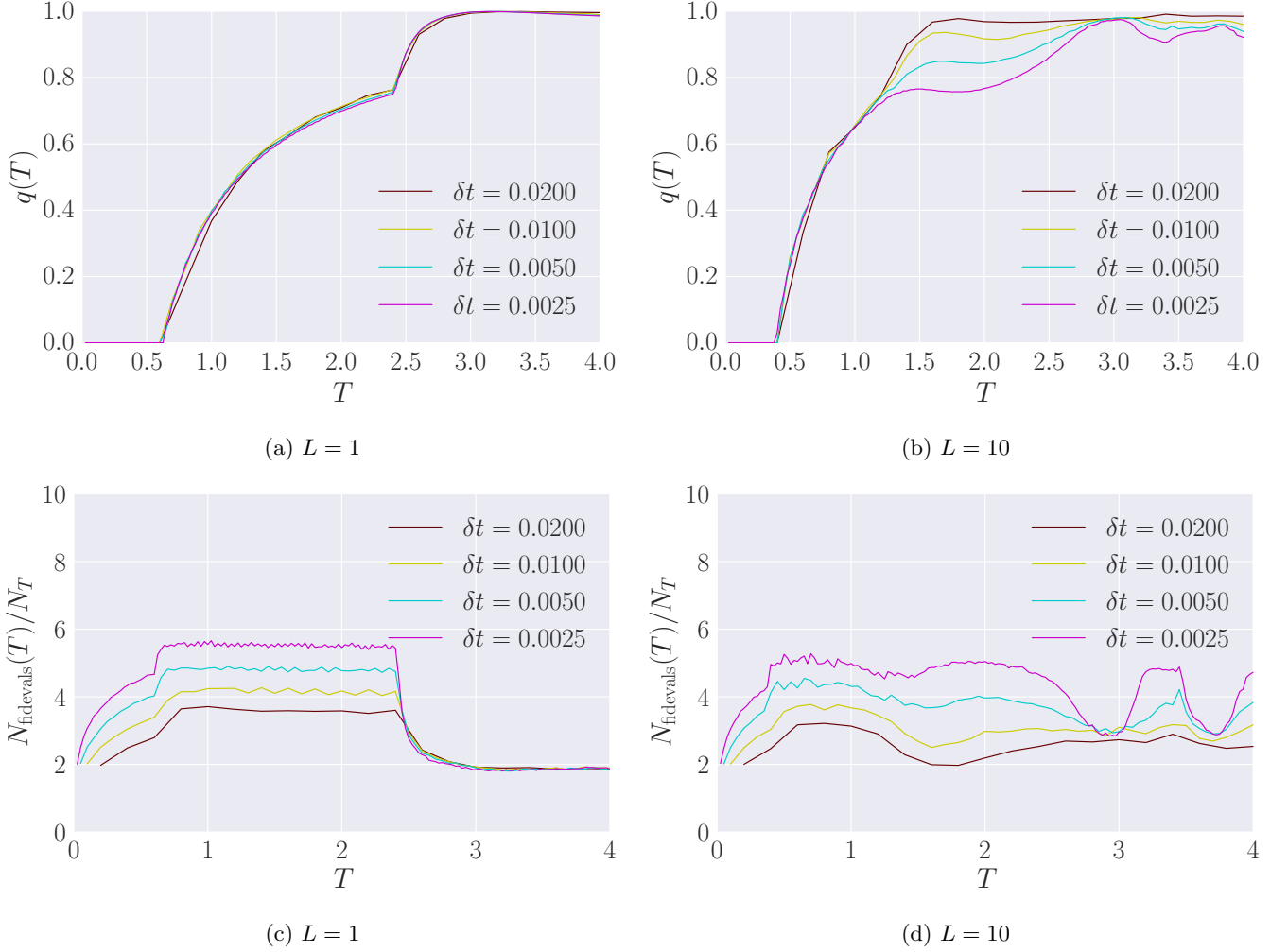
FIG. 9: Finite time step-size, δt, scaling of the order parameter $q(T)$ [top] and the number of fidelity evaluations per time step, N_{fidevals}/N_T [bottom], to reach a local minimum of the infidelity landscape.

A. Single Qubit

For $T > T_{\text{QSL}}$, the optimization problem of the single qubit ($L = 1$) is controllable, and there exist infinitely many protocols which can prepare the target state with unit fidelity. In analogy with the random k-SAT problem [54], we call this the *controllable* (or underconstrained) phase of the quantum control problem. Intuitively, this comes about due to the large total ramp times available which allow one to correct a set of 'wrong moves' at a later time in order to achieve a better fidelity at the end of the ramp. We have seen that both the Reinforcement Learning (RL) and Stochastic Descent (SD) agents readily and quickly learn to exploit this feature for optimal performance. In this phase, which is not present in the thermodynamic limit $L \to \infty$, there exist analytical results to compute driving protocols of unit fidelity based on the geodesic and counter-diabatic approaches [22, 23, 46]. The driving protocols $h_x^\alpha(t)$, corresponding to the minima of the infidelity landscape $I_h(T)$, are completely uncorrelated, resulting in $\overline{h_x}(t) = 0$ and, thus, $q = 1$. As $T \searrow T_{\text{QSL}}$, the infidelity minima start becoming correlated, reflected in a drop in the value of $q(T)$. At the time $T = T_{\text{QSL}}$, a phase transition occurs to a glassy phase with shallow, quasi-degenerate infidelity minima corresponding to many almost equally optimal protocols. Fig. 9a shows that the order parameter $q(T)$ develops a clear non-analyticity in the continuum limit $\delta t \to 0$, which proves the existence of a phase transition in protocol space. At the same critical time T_{QSL}, a drastic rapid increase is detected in the number of fidelity evaluations required to map out the infidelity minima, see Fig. 9c.

For $T_c < T < T_{\text{QSL}}$ the control problem is in the glassy phase. We showed in the main text by examining the DOS of protocols with respect to local flips of the field, that finding the optimal protocol appears as complicated as finding
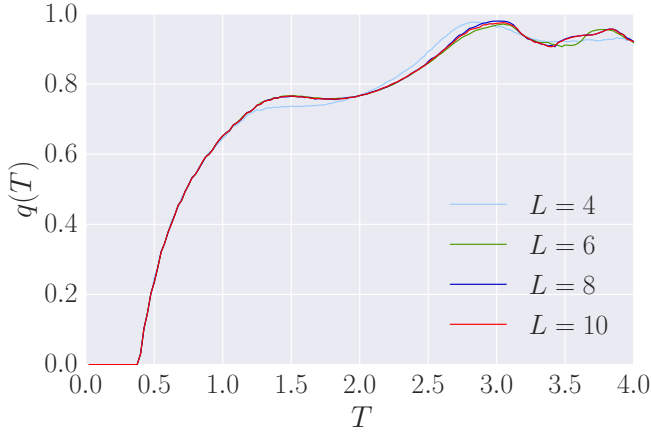
(a) Order parameter $q(T)$ vs. total ramp duration T.
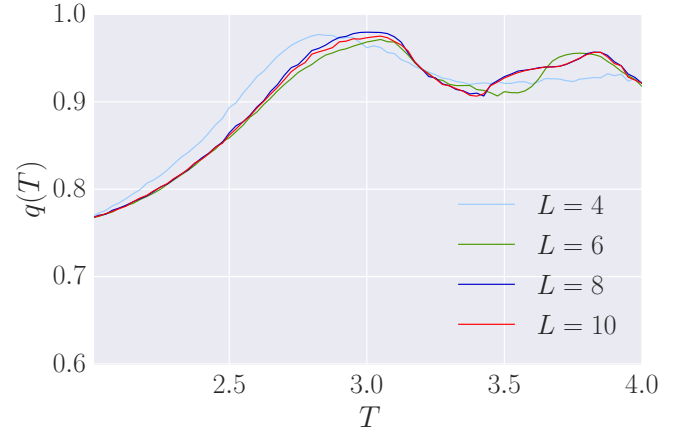
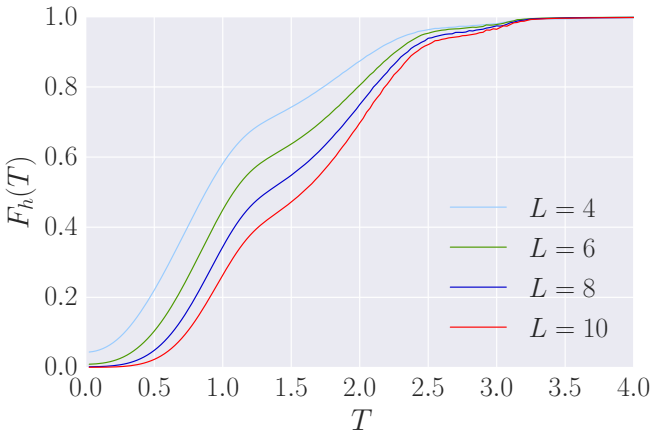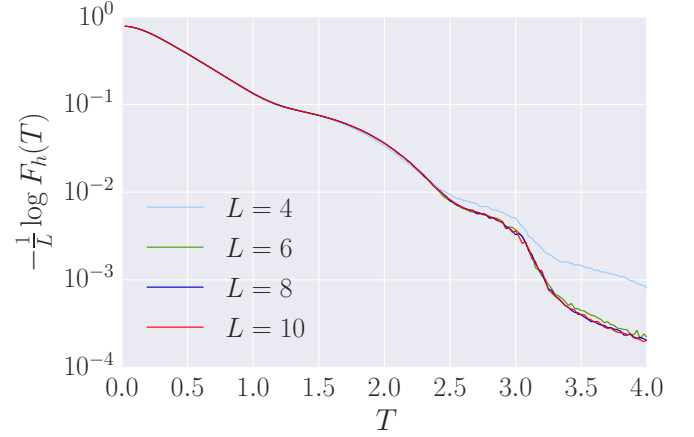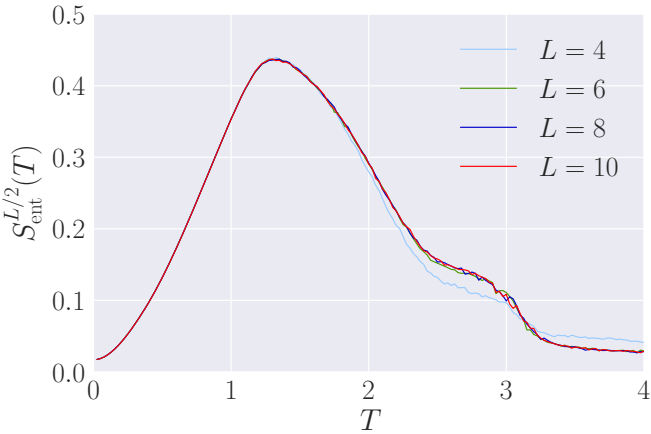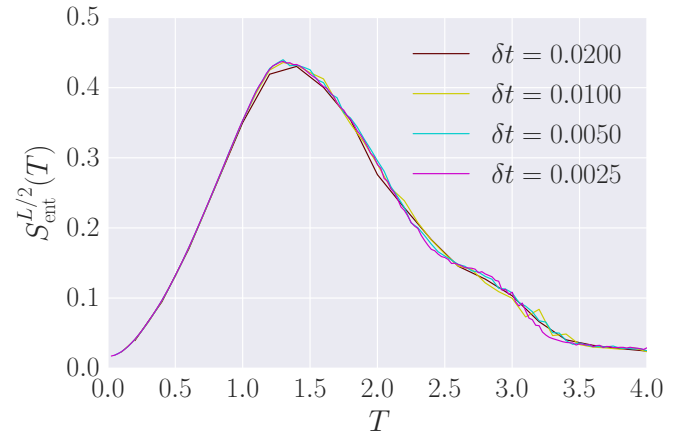(b) same as (a) with later times zoomed in.

(c) Optimal protocol many-body fidelity: linear scale. Convergence is reached at $L \geq 6$.

(d) Optimal protocol many-body fidelity: logarithmic scale. Convergence is reached at $L \geq 6$.

(e) Entanglement entropy of the half chain as a function of the system size L.

(f) Entanglement entropy of the half chain as a function of the protocol step size δt, for $L = 10$.

FIG. 10: Finite system-size L scaling of the order parameter $q(T)$[top], the many-body fidelity $F_h(T)$ [middle] and the entanglement entropy $S_{\text{ent}}^{L/2}$ for protocol step size $\delta t = 0.0025$.

the ground state of a glass. This is reflected in the observed increase of the average number of fidelity evaluations

N_{fidevals} with decreasing δt (c.f Fig. 9c), and a decrease in the learning rate of the RL and SD agents. The local minima protocols $\{h_x^\alpha(t)\}$ are strongly correlated, as can be seen from the finite value of the order parameter $q(T)$ in Fig. 10a. More importantly, for any practical purposes, unit fidelity can no longer be obtained under the given dynamical constraints.

When we reach the second critical point $T = T_c$, another phase transition occurs from the glassy to an overconstrained phase. At $T = T_c$, the order parameter reaches zero, suggesting that the infidelity landscape contains a single minimum. In this phase, i.e. for $T < T_c$, the ramp time is too short to achieve a good fidelity. Nonetheless, in the continuum limit $\delta t \to 0$, there exists a single optimal protocol, although the corresponding maximum fidelity is far from unity. In this overconstrained phase, the optimization problem becomes convex and easy to solve. This is reflected by the observation both the optimal quasi-continuous and bang-band protocols found by the RL agent are nearly identical, cf. Fig. 8. The dynamic character of the phase transition is revealed by a sudden drop in the number of fidelity evaluations N_{fidevals}.

B. Coupled Qubits

One can also ask the question what happens to the quantum control phases in the thermodynamic limit, $L \to \infty$. To this end, we repeat the calculation for a series of chain lengths L. Due to the non-integrable character of the many-body problem, we are limited to small system sizes. However, Fig. 10 shows convincing data that we capture the behaviour of the system in the limit $L \to \infty$ for the relatively short ramp times under consideration. Moreover to our surprise the finite-size effects almost entirely disappear for $L \geq 6$ for all range of ramp times we are considering. It seems that system is able to find an optimal solution, where the information simply does not propagate outside of a very small region and hence the optimal protocol rapidly becomes completely insensitive to the system size.

Figure 10c-d shows the system size scaling of the negative logarithmic *many-body* fidelity. While at $L = 4$ we do see remnants of finite-size effects, starting from $L = 6$ the curves are barely changing. A similar rapid system-size convergence is observed also for the order-parameter $q(T)$ (see Fig. 10a-b) and the entanglement entropy of the half chain (Fig. 10e). The protocol step size dependence of the order parameter $q(T)$, the average fidelity evaluations N_{fidevals}, and the entanglement entropy $S_{\text{ent}}^{L/2}$ of the half-chain are shown in Figs. 9b, 9d and 10f.

V. USING INSIGHTS FROM MACHINE LEARNING TO CONSTRUCT A VARIATIONAL THEORY FOR THE PHASE DIAGRAM OF THE QUANTUM CONTROL PROBLEM

In the main text we mentioned the possibility to use the computer agent to synthesise information into a handful effective degrees of freedom of the problem under consideration, and then use those to construct effective theories for the underlying physics. Let us now demonstrate how this works by giving a specific example which, to our surprise, captures the essence of the phase diagram of quantum control both qualitatively and quantitatively.

A. Single Qubit

We start, stealing some ideas from the computer agent, by carefully studying the optimal driving protocols it finds in the case of the single qubit. Focussing on bang-bang protocols once again, and looking at the optimal drives for the overconstrained and correlated phases, cf. Fig. 8a-b and Movies 1–3, we recognize an interesting pattern: for $T < T_c$, as we explained in the main text, there is only one minimum in the infidelity landscape, which dictates a particularly simple form for the bang-bang protocol – a single jump at half the total ramp time $T/2$. On the other hand, for $T_c \leq T \leq T_{\text{QSL}}$, there appears a sequence of multiple bangs around $T/2$, which grows with increasing the ramp time T. By looking at the Bloch sphere representation, see Movies 1–3, we identify this as an attempt to turn off the h_x-field, once the state has been rotated to the equator, so the instantaneous state can be moved in the direction of the target state in the shortest possible distance [i.e. along a geodesic].

Hence, it is suggestive to try out a three-pulse protocol as an optimal solution, see Fig. 11a: the first (positive) pulse of duration $\tau^{(1)}/2$ brings the state to the equator. Then the h_x-field is turned off for a time $\tilde{\tau}^{(1)} = T - \tau^{(1)}$, after which a negative pulse directs the state off the equator towards the target state. Since the initial value problem is time-reversal symmetric for our choice if initial and target states, the duration of the third pulse must be the same as that of the first one. We thus arrive at a variational protocol, parametrised by $\tau^{(1)}$, see Fig. 11a.

The optimal fidelity is thus approximated by the variational fidelity $\mathcal{F}_h(\tau^{(1)}, T - \tau^{(1)})$ for the trial protocol [Fig. 11a],
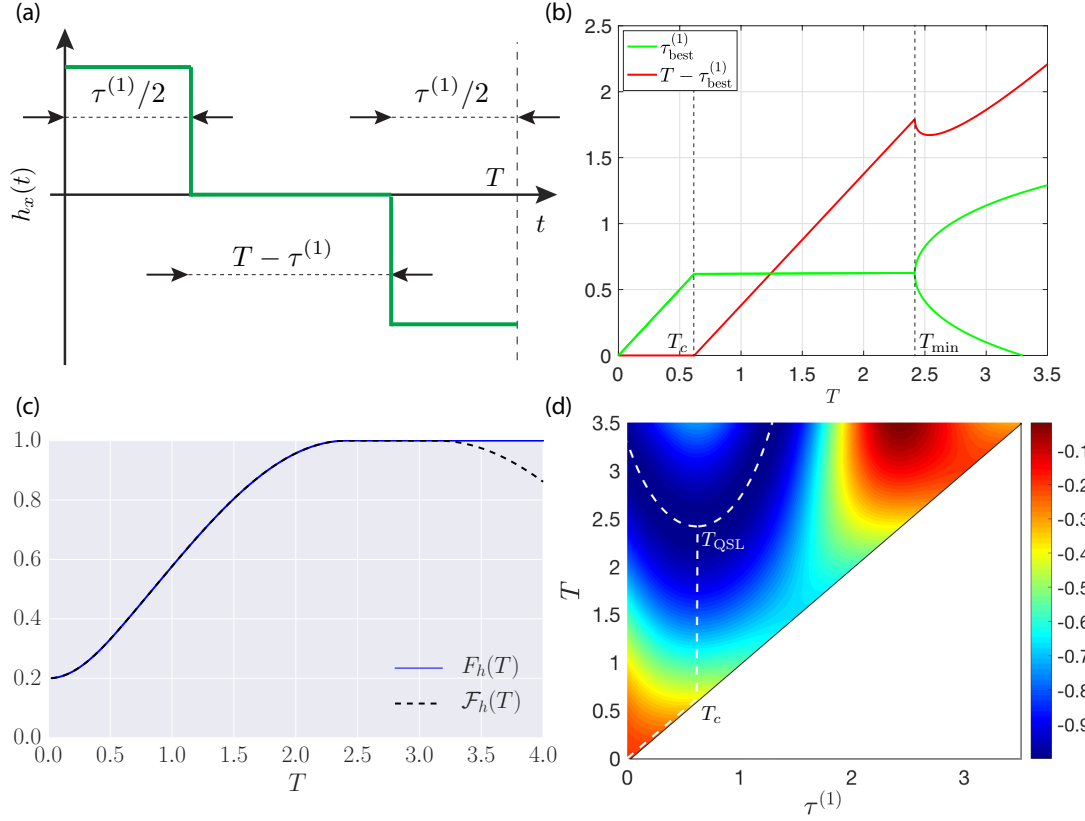
FIG. 11: (a) Three-pulse variational protocol which allows to capture the optimal protocol found by the computer in the overconstrained and the glassy phases of the single qubit problem. (b) $\tau_{\text{best}}^{(1)}$ (green), with the non-analytic points of the curve marked by dashed vertical lines corresponding to $T_c \approx 0.618$ and $T_{\text{QSL}} \approx 2.415$. (c) Best fidelity obtained using SD (solid blue) and the variational ansatz (dashed black). (d) The variational infidelity landscape with the minimum for each T-slice designated by the dashed line which shows the robustness of the variational ansatz against small perturbations.

and can be evaluated analytically in a straightforward manner:

$$\mathcal{F}_h(\tau^{(1)}, T - \tau^{(1)}) = |\langle\psi_*|e^{-i\frac{\tau^{(1)}}{2}(S^z - h_{\max}S^x)}e^{-i(T-\tau^{(1)})S^z}e^{-i\frac{\tau^{(1)}}{2}(S^z + h_{\max}S^x)}|\psi_i\rangle|^2. \tag{16}$$

The exact expression is rather cumbersome and we choose not to show it explicitly. Optimizing the variational fidelity at a fixed ramp time T, we solve the corresponding transcendental equation to find the extremal value $\tau_{\text{best}}^{(1)}$, and the corresponding optimal variational fidelity $\mathcal{F}_h(T)$, shown in Fig. 11b-c. For times $T \leq T_c$, we find $\tau^{(1)} = T$ which corresponds to $\tilde{\tau}^{(1)} = 0$, i.e. a single bang in the optimal protocol. The overconstrained-to-glassy phase transition at T_c is marked by a non-analyticity at $\tau_{\text{best}}^{(1)}(T_c) = T_c \approx 0.618$. This is precisely the minimal time the agent can take, to bring the state to the equator of the Bloch sphere, and it depends on the value of the maximum magnetic field allowed (here $h_{\max} = 4$). Figure 11d shows that, in the overconstrained phase, the fidelity is optimised at the boundary on the variational domain, although $\mathcal{F}_h(\tau^{(1)}, T - \tau^{(1)})$ is a highly nonlinear function of $\tau^{(1)}$ and T.

For $T_c \leq T \leq T_{\text{QSL}}$, the time $\tau^{(1)}$ is kept fixed (the equator being the only geodesic for a rotation along the $\hat{z}$-axis of the Bloch sphere), while the second pulse time $\tilde{\tau}^{(1)}$ grows linearly, until the minimum time $T_{\text{QSL}} \approx 2.415$ is eventually reached. The minimum time is characterised by a bifurcation in our effective variational theory, as the corresponding variational infidelity landscape develops two maxima, see Fig. 11b,d. Past that point, our simplified ansatz is no longer valid, and the system is in the controllable phase. Furthermore, a sophisticated analytic argument based on optimal control theory can give exact expressions for T_c and T_{QSL} [46].

B. Coupled Qubits

Let us also discuss the variational theory for the many-body system. Once again, inspired by the structure of the protocols found by the computer, see Movie, we extend the qubit variational protocol, as shown in Fig. 12c. In particular, we add two more pulses to the protocol, retaining its symmetry structure: $h_x(t) \longrightarrow -h_x(T - t)$, whose length is parametrised by a second, independent variational parameter $\tau^{(2)}/2$. Thus, the pulse length where the field is set to vanish, is now given by $\tilde{\tau} = T - \tau^{(1)} - \tau^{(2)}$. This pulses are reminiscent of spin-echo protocols, and appear to be important for entangling and disentangling the state during the evolution.
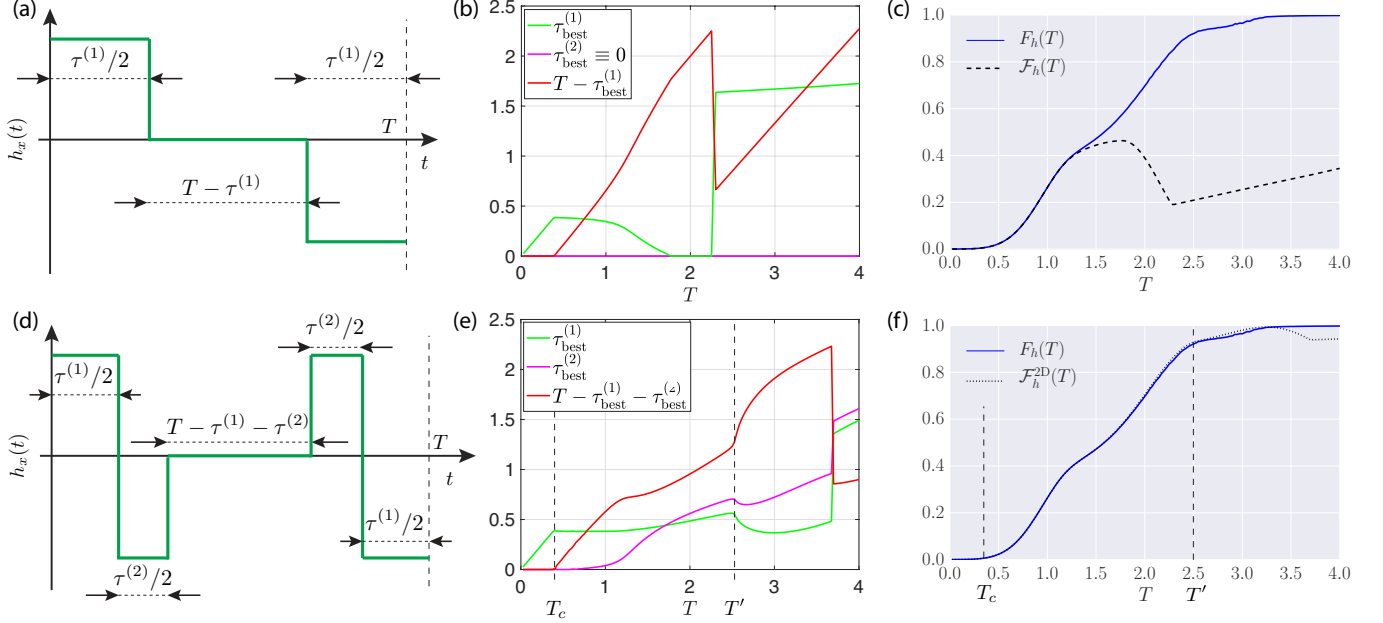

FIG. 12: (a) Three-pulse variational protocol which allows to capture the optimal protocol found by the computer in the overconstrained phase but fails the glassy phase of the nonintegrable many-body problem. This ansatz captures the non-analytic point at $T_c \approx 0.4$ but fails in the glassy phase. (b) The pulse durations $\tau^{(1)}_{\text{best}}$ (green) and $\tau^{(2)}_{\text{best}}$ (magenta), for highest fidelity variational protocol of length T of the type shown in (a). The fidelity of the variational protocols exhibit a physical non-analyticity at $T_c \approx 0.4$ and unphysical kinks outside the validity of the ansatz. (c) 1D maximal variational fidelity (dashed back) compared to the best numerical protocol (solid blue). (d) Five-pulse variational protocol which allows to capture the optimal protocol found by the computer in the overconstrained phase and parts of the glassy phase of the nonintegrable many-body problem. (e) The pulse durations $\tau^{(1)}_{\text{best}}$ (green) and $\tau^{(2)}_{\text{best}}$ (magenta) for the best variational protocol of length T of the type shown in (d). These variational protocols exhibit physical non-analyticities at $T_c \approx 0.4$ and $T' \approx 2.5$ (vertical dashed lines) (f) 2D maximal variational fidelity (dashed-dotted back) compared to the best numerical protocol (solid blue).

Notice that this extended variational ansatz includes by definition the simpler ansatz from the single qubit problem discussed above, by setting $\tau^{(2)} = 0$. Let us focus on this simpler case first. The dashed black line in Fig. 12c shows the corresponding 1D variational fidelity. We see that, once again this ansatz captures correctly the critical point T_c separating the overconstrained and the glassy phases. Nevertheless, a comparison with the optimal fidelity [see Fig. 12c] reveals that this variational ansatz breaks down in the glassy phase, although it rapidly converges to the optimal fidelity with decreasing T for $T < T_c$. Going back to Fig. 12b, we note that the value $\tau^{(1)}_{\text{best}}$ which maximizes the variational fidelity exhibits a few kinks. However, only the kink at $T = T_c$ captures a physical transition of the original control problem, while the others appear as artefacts of the simplified variational theory.

Let us now turn on the second variational parameter $\tau^{(2)}$, and consider the full two-dimensional variational problem:

$$\mathcal{F}_h^{2D}(\tau^{(1)}, \tau^{(2)}, T - \tau^{(1)} - \tau^{(2)}) = \tag{17}$$

$$|\langle \psi_* | e^{-i\frac{\tau^{(1)}}{2}(S^z - h_{\max}S^x)} e^{-i\frac{\tau^{(2)}}{2}(S^z + h_{\max}S^x)} e^{-i(T - \tau^{(1)})S^z} e^{-i\frac{\tau^{(2)}}{2}(S^z - h_{\max}S^x)} e^{-i\frac{\tau^{(1)}}{2}(S^z + h_{\max}S^x)} |\psi_i\rangle|^2.$$

Figure 12 shows the best variational fidelity $\mathcal{F}_h^{2D}$ [f] and the corresponding values of $\tau^{(1)}_{\text{best}}$ and $\tau^{(2)}_{\text{best}}$ [e] for this maximum fidelity variational protocol [d]. There are two important points here: (i) Fig. 12f shows that the 2D variational fidelity

seemingly reproduces the optimal fidelity on a much longer scale, i.e. for all protocol durations $T \lesssim 3.3$. (ii) the 2D variational ansatz reduces to the 1D one in the overconstrained phase $T \leq T_c$. In particular, both pulse lengths $\tau_{\text{best}}^{(1)}$ and $\tau_{\text{best}}^{(2)}$ exhibit a non-analyticity at $T = T_c$, but also at $T' \approx 2.5$. Interestingly, the 2D variational ansatz captures the optimal fidelity on both sides of T' which suggests that there is likely yet another transition within the glass phase, hence the different shading in the many-body phase diagram [Fig. 2, main text]. Similar to the 1D variational problem, here we also find artefact transitions (non-analytic behavior in $\tau_{\text{max}}^{(i)}$) outside of the validity of the approximation.

In summary, we have shown how, by carefully studying the driving protocols found by the computer agent, one can obtain ideas for effective theories which capture the essence of the underlying physics. This is similar to using a ϕ^4-theory to describe the physics of the Ising phase transition. The key difference is that the present problem is out-of-equilibrium, where no general theory of statistical mechanics exists so far. We hope that, an underlying pattern between such effective theories can be revealed with time, which might help shape the principles of a theory of statistical physics away from equilibrium.