# **Shaping Laser Pulses with Reinforcement Learning**

Anonymous authors Paper under double-blind review

Keywords: Applied RL, DRL for Science, Sim-to-real, Domain Randomization, Ultra-short pulses

# **Summary**

High Power Laser (HPL) systems operate in the femtosecond regime—one of the shortest timescales achievable in experimental physics. These systems are instrumental in high-energy physics, as ultra-short pulses yield extremely high intensities, which play out to be essential in both practical applications and theoretical advancements in light-matter interactions. Traditionally, the parameters regulating HPL optical performance are optimized using black-box numerical methods such as Evolution Strategies (ES), and Bayesian Optimization (BO). While effective, black-box methods are computationally demanding and rely on stationarity assumptions overlooking transient and complex system dynamics in HPL systems. Moreover, their safe implementation on real-world hardware is challenging, as erratic exploration of the parameter space can compromise system safety. Model-free Deep Reinforcement Learning (DRL) offers a promising alternative by enabling sequential decision making in non-static settings. This work investigates the safe application of DRL to HPL systems, and extends current research by (1) learning a control policy directly from pixels, using images typically available in experimental settings and (2) addressing the need for generalization across diverse dynamics, tackling the non-stationarity of the environment. We evaluate our method in simulation across various dynamic configurations and observe that DRL effectively enables cross-domain adaptability by transferring knowledge across conditions—eliminating the need to restart ES/BO whenever there are fluctuations in the environment dynamics. Our contributions represent a significant step towards real-world applications of DRL to HPL systems, introducing the RL community to the task of controlling complex non-linear physical systems used to ignite nuclear fusion and accelerate charged particles.

# **Contribution(s)**

- We demonstrate the benefits of using Deep Reinforcement Learning to optimize High Power Laser systems over currently dominant approaches based on gradient-free optimization.
   Context: Prior works on laser optimization focused on black-box optimization techniques which assume stationarity, require costly real-world function evaluations, and can endanger the system at test-time.
- We learn a control policy directly from images, which are made available via widespread diagnostics devices.

**Context:** Instead of relying on noisy and lengthy processes to obtain structured representations of the system's state, we leverage unstructured observations coming from diagnostic devices as inputs for the control policy.

3. We train control policies entirely in simulation and successfully transfer them across unknown, varying dynamics, showing robustness to different parametrizations. Context: Transferring policies is hindered by the discrepancies across different domains, and Domain Randomization is a promising technique widely explored in the field of robot learning to ensure robustness to said differences and thus overcome such limitation.

# **Shaping Laser Pulses with Reinforcement Learning**

#### **Anonymous authors**

Paper under double-blind review

#### Abstract

1 High Power Laser (HPL) systems operate in the femtosecond regime-the shortest 2 timescale achievable in experimental physics. HPL systems are instrumental in high-3 energy physics, leveraging ultra-short impulse durations to yield extremely high inten-4 sities, which are essential for both practical applications and theoretical advancements 5 in light-matter interactions. Traditionally, the parameters regulating HPL optical perfor-6 mance are tuned manually by human experts, or optimized by using black-box methods 7 that can be computationally demanding. Critically, black box methods rely on stationar-8 ity assumptions overlooking complex dynamics in high-energy physics and day-to-day 9 changes in real-world experimental settings, and thus need to be often restarted. Deep Reinforcement Learning (DRL) offers a promising alternative by enabling sequential 10 11 decision making in non-static settings. This work investigates the safe application of 12 DRL to HPL systems, and extends the current research by (1) learning a control policy 13 directly from images and (2) addressing the need for generalization across diverse dy-14 namics. We evaluate our method across various configurations and observe that DRL 15 effectively enables cross-domain adaptability, coping with dynamics' fluctuations while 16 achieving 90% of the target intensity in test environments.

## 17 1 Introduction

Ultra-fast light-matter interactions find applications in both theoretical and experimental physics. The extremely high intensities—in the order of petawatts—that can be attained with modern-day High Power Laser (HPL) systems enable a variety of use cases in light-matter interactions and charged-particles acceleration. Extreme intensities are typically achieved by focusing high-energy laser pulses onto spatial targets for ultra-short durations—down to attoseconds. As a result, ultrashort laser pulses represent the shortest events ever created by humanity (Gaumnitz et al., 2017).

24 Over the course of 2022 and 2023, four separate experiments at the Lawrence Livermore National 25 Laboratory (LLNL)-National Ignition Facility (USA) employed HPL systems to achieve nuclear fu-26 sion ignition (Abu-Shawareb et al., 2024). In their experiments, the scientists at the LLNL used 192 27 HPL beams to achieve nuclear fusion ignition in a laboratory setting, and went on demonstrating 28 larger-than-unity energy gains, achieving energy-positive results in nuclear fusion. HPL systems 29 also have applications in radiation-based cancer therapy, as they can be used to produce beams of 30 high-energy charged particles, which interact with malignant cells and thus yield radio-therapeutic 31 outcomes (Grittani et al., 2020). Lastly, HPL systems enable the controlled study of the interac-32 tion between extremely intense beams of light and various materials, providing valuable insights to 33 numerous scientific communities, including plasma, laser and theoretical physicists.

HPL systems' performance heavily depend on environmental conditions, and on numerous parameters. For instance, HPL systems are typically operated in remote areas or meters underground to mitigate road-induced vibrations that might cause misalignment in the optics. Further, HPL systems are run in environmentally controlled facilities (*cleanrooms*), to prevent airborne particles to sediment on the optical gear. Parameters-wise, *dispersion coefficients* play a central role, as they physically determine the phase shifts imposed on the different frequencies of the light beam. In



Figure 1: (A) Schematic representation of the RL pipeline for pulse shaping in HPL systems. The model processes images to produce phase corrections, leading to shorter pulse durations and intensity maximization. To improve on robustness, during training the agent faces a distribution of dynamics rather than a single one. (B) Illustration of the process of linear and non-linear phase accumulation taking place along the pump-chain of HPL systems. By opportunely controlling the phase imposed at the stretcher, one can benefit from both energy and duration gains, for maximal peak intensity.

40 turn, this leads to shorter laser pulses and intensity gains when the applied phase induces construc-

41 tive interferences between frequencies, whereas destructive interference results in longer pulses and 42 intensity losses (Paschetta, 2008)

42 intensity losses (Paschotta, 2008).

43 Traditionally, laser parameters have been optimized using 1D searches over the range of possible values. More recently, black-box numerical methods such as Evolution Strategies (ES) and Bayesian 44 45 Optimization (BO) have been studied (Loughran et al., 2023; Shalloo et al., 2020; Arteaga-Sierra et al., 2014). While effective, these black-box methods can be computationally demanding, as they 46 47 are typically implemented on real-world laser systems, and thus require costly laser-bursts to per-48 form one single function evaluation. Further, they rely on stationarity assumptions overlooking 49 transient and complex non-linear system dynamics. Lastly, their safe implementation on real-world 50 hardware can be challenging, as erratic exploration of the parameter space can compromise system 51 safety (Capuano et al., 2023).

This work investigates the safe application of DRL to HPL systems for temporal profile shaping 52 53 via autonomous, bounded control of the dispersion coefficients. In particular, we present an ap-54 plication of DRL to intensity maximization through pulse duration minimization. We leverage an 55 openly-available simulator (Capuano et al., 2023) of a component of the world's most powerful laser 56 system, and learn an adaptive control policy capable of safely tuning the dispersion coefficients for 57 intensity maximization. In our work, we simulate different experimental conditions by arbitrarily 58 randomizing parameters of our simulator, and use said randomization over the laser system dynam-59 ics to induce the learned policy to be robust to changes in the experimental setting (Tiboni et al., 60 2023c). As parameters of HPL systems can typically only be estimated and vary over time, robust-61 ness is paramount for a wide applicability of our approach. To further improve on this and pave 62 the way towards real-world applications of RL to HPL systems, we also leverage Deep Learning 63 to process unstructured observations in the form of readily available images (FROG traces). Our 64 contributions can be summarized as follows:

We present an application of DRL to the rich and complex domain of experimental laser
 physics, demonstrating its suitability for handling the non-stationarity and transient non-linear
 dynamics of HPL systems—challenges often overlooked by predominant black-box approaches.

• We **train control policies entirely in simulation and successfully transfer them** across different environments, ensuring adaptivity to (1) inaccuracies in parameter estimation and (2) evolution of experimental setting. Randomizing also helps mitigate the impact due to under-modeled dynamics in simulation. 72 • We learn a control policy from single-channel images readily available in most experimen-

tal settings, using them as a proxy for pulse duration. This eliminates the need for quantumdestructive measurements on charged particles' energy, or noisy temporal pulse reconstruction.

destructive measurements on charged particles' energy, or noisy temporal pulse reconstruction,
 and enables a real-time feedback loop using existing experimental hardware—making our method

76 more applicable in real-world settings.

## 77 2 Background & Related Work

#### 78 2.1 Optimizing Laser Systems

79 Traditionally, HPL systems' parameters have been optimized using independent 1D grid-searches 80 over all the considered dimensions. While straightforward, this approach naively overlooks the joint 81 effect varying multiple parameters simultaneously can have on the system. More recently, Evolution 82 Strategies (ES) (Baumert et al., 1997; Arteaga-Sierra et al., 2014; Woodward & Kelleher, 2016), and Bayesian Optimization (BO) (Loughran et al., 2023; Shalloo et al., 2020; Capuano et al., 2022; An-83 84 jum et al., 2024) have been proposed to optimize HPL performance. Differently from grid-search, 85 ES and BO do take into account the joint effect of different parameters on the system, and proved 86 effective real-world experiments (Shalloo et al., 2020). However, while performant, black-box meth-87 ods tend to be computationally demanding in the number of functions evaluations—real-world laser 88 bursts—and typically do not provide guarantees regarding the stability of the control configuration 89 found to changes in the environment. That is, for any changes in the experimental condition one 90 could need to re-optimize the system from scratch, just as humans do. Further, these algorithms 91 rely on stationarity assumptions within experimental conditions, overlooking the transient and com-92 plex dynamics characteristic of high-intensity phase accumulation processes in non-linear crystals. 93 Lastly-differently from grid search-the safe implementation of black-box methods on real-world 94 hardware can be challenging, as gains in sample efficiency might trade-offs with erratic exploration 95 of the parameter space (Capuano et al., 2023), endangering system's safety.

96 To allow for a more adaptive control of laser systems, recent works have investigated the application 97 of Reinforcement Learning (RL) to HPL systems (Kuprikov et al., 2022; Rakhmatulin et al., 2024; 98 Mareev et al., 2023; Capuano et al., 2023). Mareev et al. (2023) investigated the application of 99 DRL to maintain a laser beam focused on a solid target, shifting away as a consequence of high-100 energy light-matter interactions and thus requiring constant target-position adjustment. Rakhmatulin 101 et al. (2024) investigated the application of RL to the problem of optics alignment in laser systems, 102 controlling the position of mirrors via real-time camera feedback. While both target location and 103 mirror alignment have a significant impact on the final intensity conveyed by the beam, neither directly shapes the temporal profile of laser pulse and thus the final peak intensity. Kuprikov et al. 104 105 (2022) learned a controller to adaptively adjust the power supplied to the laser, and the filters used to 106 temporally shape the output, thus directly impacting peak intensity. However, the authors considered 107 the problem of ensuring highly-similar pulses between multiple laser bursts, by learning to mode-108 lock the system, rather than shaping the individual pulse to be obtained. Capuano et al. (2023) 109 studies the problem of learning a controller for pulse shaping, by directly tuning the dispersion 110 coefficients and thus ensuring a closer loop between control parameters and peak intensity. However, 111 in their work Capuano et al. (2023) overlook several practical aspects associated with deploying 112 control policies to real world laser systems, such as the necessity of coping with possibly imprecise 113 estimates of the experimental setting, and the need to adapt to the non-stationary of the experimental 114 environment. Unlike previous attempts at temporal pulse shaping, we work backwards from real-115 world deployment requirements, extending the current research by learning a robust control policy 116 for the dispersion coefficients that is (1) machine-safe to deploy, (2) inherently adaptive and (3) uses 117 readily available information in most HPL diagnostic systems.

#### 118 2.2 Shaping Laser Pulses

119 The optimization of laser pulse shape and duration is a critical challenge in HPL systems, par-120 ticularly for applications in laser-plasma acceleration, high-intensity laser-matter interactions, and 121 inertial confinement fusion. Furthermore, the precise control of pulse shape directly influences 122 the peak intensity, energy deposition efficiency, and nonlinear optical effects encountered during 123 the laser propagation itself. In applications of HPL systems to charged particle acceleration (Grit-124 tani et al., 2020), directly measuring the particles' beam energy is a quantum-destructive process— 125 charged particles lose their energy when an experimental energy probe interacts with them. How-126 ever, proxying particles' beam energy with pulse's peak intensity, HPL systems can be optimized 127 using the peak intensity  $I^*$  produced. At iso-energy, intensity maximization takes place by min-128 imizing the pulse duration, measured by its full-width half-maximum (FWHM) value—the value  $|t_l - t_r| : I(t_l) = I(t_r) = \frac{1}{2}I^*$ . Ultra-short pulses' duration is typically inferred from frequency-129 resolved optical gating (FROG) traces (Trebino & Kane, 1993), for the scope of this work considered 130 131 as single-channel *images* showing the spectral phase accumulated by a pulse. Thus, black-box meth-132 ods and 1D-grid search are fundamentally ill-posed to use these non-destructive measurements of particle beam's energy as their input, while DRL can instead fully leverage the advancements made 133 134 in Deep Learning to handle unstructured data formats as control inputs (Mnih et al., 2013).

135 In practice, HPL systems rely on the transferring of energy from a high-power primary pump laser 136 beam to a secondary *seed* laser beam. The spectral and temporal characteristics of the pump laser 137 determine much of the achievable pulse intensity. Critically, for the sake of intensity gains in the 138 seed laser, the pump laser is usually run through an amplification chain introducing both linear 139 and nonlinear phase distortions. As phase regulates how the spectral intensity overlays in the time 140 domain (Paschotta, 2008), it must be carefully controlled to achieve efficient amplification at the 141 pump and seed level. Typically, pump chains follow a Chirped Pulse Amplification (CPA) scheme. 142 Figure 1 illustrates the CPA process, where the initial pump pulse is (1) stretched in time to avoid 143 nonlinear effects and damage to the earlier stages of the pump chain due to high intensities (2) 144 amplified via regenerative and multipass amplifiers, and (3) re-compressed in time to achieve high 145 peak intensity.

146 Unlike the amplification and compression stages, the process of pulse stretching can typically be 147 controlled externally from laser specialists, varying the dispersion coefficients of the phase of the pump laser applied. The spectral phase of a laser beam  $\varphi(\omega)$  is typically modeled using a Taylor 148 expansion around the central angular frequency of the pulse  $\omega_0$ , yielding  $\varphi(\omega) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{\partial^k \varphi}{\partial \omega^k} (\omega - \omega_0)^k$ . The first two terms in this polynomial expansion— $\varphi(\omega)$  and  $\varphi'(\omega)(\omega - \omega_0)$ —do not directly 149 150 151 influence the shape of the pulse in the temporal domain. Conversely, second-order (group-delay 152 dispersion, GDD), third-order (third-order dispersion, TOD) and fourth-order (fourth-order disper-153 sion, FOD) derivatives—jointly referred to with  $\psi = (\text{GDD}, \text{TOD}, \text{FOD}) \in \Psi$ —do influence the 154 resulting temporal profile. By opportunely tuning  $\psi$ , laser specialists are able to control the tempo-155 ral profile of ultra-short laser pulses. Physically, control over  $\psi$  is achieved using a Chirped Fiber 156 Bragg Grating (CFBG), consisting of an optical fiber whose grating is adjusted inducing a tempera-157 ture gradient at its extremes. Consequently, it is crucial to carefully regulate the relative temperature 158 variations to avoid demanding abrupt control adjustments over short time intervals, which could 159 damage the fiber.

160 In the context of laser optimization, one might want to maximize the intensity conveyed by a laser 161 pulse by minimizing its duration, i.e. performing *temporal shaping* by controlling  $\psi$ . Typically, 162 highly trained human experts spend hours carefully varying  $\psi$  in the real world, leveraging a mix of 163 past experience and personal expertise at the task. The shortest time duration attainable by a laser 164 pulse is typically referred to as Transform Limited (TL), and corresponds to perfect overlay of all 165 the different spectral components of intensity in time-as such, it has an accumulated phase equal 166 to  $\varphi^*(\omega) = 0$ . Critically, the amplification step in CPA introduces nonlinear phase components. If 167 this was not the case, then one could retrieve TL pulses by simply applying a phase at the stretcher level that is opposite to the one defined at the compressor's,  $\varphi_s(\omega) = -\varphi_c(\omega)$ . However, the non-168

169 linearity induced by the amplification step calls for a more sophisticated control over  $\varphi_c(\omega)$ . This

difficulty arises from the need to balance non-linear effects in the phase accumulation process and non-stationary experimental conditions, while adhering to a sequential control approach that ensures

171 non-stationary experimental conditions, while adhering to a sequential control approach that ensures

machine safety by limiting abrupt changes in control parameters.

#### 173 2.3 Sim-to-real

Even the most sample efficient of the numerical algorithms typically considered for pulse shaping varying dispersion coefficients can require 10<sup>2</sup> samples (Capuano et al., 2022), corresponding to just as many real-world laser bursts (Shalloo et al., 2020). Such computational demands are hard to meet in real-world systems, and are especially more troubling if one considers the instability of the solution found with respect to changes in the experimental setting. Further, BO can endanger the system by applying abrupt controls at initialization.

180 We can mitigate the need for expensive real-world data samples by leveraging simulated versions of 181 the phase accumulation process, where we can easily accommodate for large number of samples, as 182 well as safe exploration of the dispersion coefficients space,  $\Psi$ . While typically not accurate enough to directly transfer point-solutions  $\psi^*$  from simulations to the real world, simulators can be used 183 184 to train control policies for different environments. The problem of transferring control policies 185 across domains is a well-studied problem in applications of RL for robotics, and the community has 186 extensively investigated approaches to crossing the *reality gap* (Tobin et al., 2017; Valassakis et al., 187 2020). Considering this last point, we argue the HPL setting closely resembles the challenges the 188 community faces when transferring robotic policies across environments.

Transferring a control policy across diverse environments can be achieved (1) reducing the discrepancy between them (Zhu et al., 2017) or (2) applying parameter randomization to improve on the robustness of the policy (Peng et al., 2018). One widely adopted sim-to-real method is Domain Randomization (DR), which involves varying simulator parameters within a predefined distribution during training (Valassakis et al., 2020) to incentivize generalization over said parameters. DR introduces additional sources of stochasticity into the environment dynamics, making policies more robust at an increased risk of sub-optimality and over-regularization (Margolis et al., 2024).

196 Although having proved effective on robotics tasks (Antonova et al., 2017), DR suffers from the key 197 limitation of needing to extensively tune the distributions used in training. Automated approaches 198 to DR propose adaptive distribution refinement over training, e.g. by leveraging a limited set of real-world data Tiboni et al. (2023a;b), or based on the policy's performance under a given set of dy-199 200 namics parameters (Akkaya et al., 2019). While effective for dexterous manipulation, Akkaya et al. 201 (2019) has been observed to be sample inefficient, as it biases the policy towards learning dynamics 202 sampled from the boundaries of the current distribution (Tiboni et al., 2023c). A more principled 203 approach to automated DR has been recently introduced in Tiboni et al. (2023c), where the authors 204 follow the principle of maximum entropy (Jaynes, 1957) to resolve the ambiguity in defining DR 205 distributions. Particularly, the authors train adaptive control policies for progressively more diverse 206 dynamics that satisfy an arbitrary performance lower bound. Notably, the domain randomization 207 approaches in Akkaya et al. (2019); Tiboni et al. (2023c) employ history-based policies to promote 208 implicit meta-learning strategies at test time—i.e., on-line system identification.

## 209 3 Method

#### 210 3.1 MDPs for Intensity Maximization

In Capuano et al. (2023), the authors formulate pulse shaping as a control problem in a Markov Decision Process (MDP),  $\mathcal{M}$ . In this work, we extend their formulation to the case where the environment dynamics are influenced by an unobserved latent variable, leading to a *Latent MDP* (LMDP) (Chen et al., 2021), denoted as  $\mathcal{M}_{\xi} = \{S, \mathcal{A}, \mathbb{P}_{\xi}, r, \rho, \gamma\}$ . Here,  $\xi$  is a realization of a latent random vector  $\Xi$ , such that  $\xi \sim \Xi : \operatorname{supp}(\Xi) \subseteq \mathbb{R}^{|\xi|}$ , parametrizing the transition dynamics

 $\mathbb{P}_{\xi}$ . Crucially, the agent does not directly observe  $\xi$  at test time (i.e. the real world). Conversely, 216 217 we assume that parameters  $\xi$  may be accessed when training in simulation. We argue the LMDP 218 framework is particularly well-suited for pulse shaping in a non-stationary setting due to the pres-219 ence of hidden variations in the system's dynamics. In practical scenarios, an agent must adapt to 220 an unknown experimental condition which can be modeled as  $\xi$ , while iteratively refining its control 221  $\psi$ . As  $\psi$  is physically translated into temperature gradients applied to an optical fiber, the choice 222 of  $\psi_t$  must account for past applied controls, particularly  $\psi_{t-1}$ , to prevent excessive one-step tem-223 perature variations. Moreover, the day-to-day fluctuations in HPL systems can be captured through 224  $\Xi$ , modeling the inherent non-stationarity of experimental conditions. Further, by incorporating a 225 distribution over the starting condition of the system,  $\psi_0 \sim \rho$ , the pulse shaping problem's sequen-226 tial nature becomes evident—starting from a randomly sampled experimental condition, the agent must iteratively apply controls  $\psi$  while dealing with incomplete knowledge of the system dynamics. 227 228 Inspired by the domain randomization and meta-learning literatures, we therefore aim at learning 229 control policies that are robust and adaptive to unknown, hidden contexts.

230 **State space** (S) Ideally, one could access the temporal profile of the pulse to describe the status of 231 the laser system. Indeed, the temporal profile  $\chi(\psi)$  contains all the information needed to maximize 232 peak intensity, including pulse energy and duration. However, obtaining high-fidelity temporal pro-233 files of ultra-short laser pulses in practice is a challenging task (Trebino & Kane, 1993; Trebino et al., 234 1997). Here, we instead leverage FROG traces as proxy for state information. As FROG traces con-235 tain enough information to reconstruct temporal profiles (Zahavy et al., 2018), we argue they could 236 also be used as direct inputs to a control policy aiming at maximizing peak intensity. Further, using 237 FROG traces would be practically convenient given the availability of FROG detection devices in 238 most HPL systems, and prevent the need for an intermediate step in the pulse shaping feedback loop 239 to reconstruct  $\chi$  from its associated FROG trace,  $\Phi$ . Hence, we directly include FROG traces  $\Phi_t$ 240 in our state space. We complement states  $s_t$  with the vector of dispersion coefficients  $\psi_t$  and the action taken in the previous timestep,  $a_{t-1}$ , giving  $s_t = \{\Phi_t, \psi_t, a_{t-1}\}$ , as they all are information 241 242 available at test time.

Action space ( $\mathcal{A}$ ) As we are concerned with real world applicability of our method, we design an action space that is inherently machine-safe, and that can prevent erratically changing the control applied at test time. In this, we consider varying dispersion coefficients within predetermined boundaries defined at the level of the grated optical fiber, i.e.  $\psi_t \in [\psi_{\min}, \psi_{\max}] : c = |\psi_{\min} - \psi_{\max}|$ . Actions are then defined as  $a_t \in [-\alpha c, +\alpha c]$ , with  $\alpha$  being an arbitrary fraction of the total nominal range c. In our method, we set  $\alpha = 0.1$ , thus never changing  $\psi$  in one step by more than 10% of the total possible variation.

**Environment dynamics**  $(\mathbb{P}_{\xi} : S \times A \times S \mapsto \mathbb{R}^+)$  Inspired by the successes of in-simulation 250 learning in robotics (Antonova et al., 2017; Akkaya et al., 2019; Tiboni et al., 2023c), we employ 251 252 simulations of the pump chain process while training a policy to control it. This allows us to scale 253 the number of samples available at training time to amounts that are simply unfeasible on real-world 254 laser hardware. We provide a detailed description of the phase accumulation process in ??, describ-255 ing the model for state-action-next transitions,  $\mathbb{P}_{\xi}(s_{t+1}|s_t, a_t)$ . Here, we wish to pose particular emphasis on the role of  $\xi$  on  $\mathbb{P}_{\xi}$ . Figure 2 shows how different  $\xi_i$  can lead to significantly different 256 257 pulses when applying the same  $\psi$ . In particular, 2 simulates the impact of randomizing the parameter 258 regulating non-linear phase accumulation during amplification. This parameter is typically referred 259 to as *B-integral*, and indicated with *B*. In HPL systems, one cannot typically assume to have con-260 trol over B but indirectly: non-linear effects become more evident when higher-intensity pulses are 261 propagated through non-linear crystal, which induces non-stationarity in B. Further, precisely es-262 timating B at a given time is a challenging tasks, prone to imprecision and which can have drastic 263 impacts on the peak intensity achieved (Figure 3).

**Reward function** r, **Starting condition**  $\rho$  **and discount factor**  $\gamma$  We exploit our knowledge of HPL systems to design a reward function defined as the ratio between the current-pulse peak



Figure 2: Impact of the B-integral parameter on the temporal profile (top) and FROG trace (bottom).

Figure 3: Impact of longer pulses on the peak intensity conveyed, measured as a fraction of  $I_{TL}$ .

intensity  $I_t^*$  and the highest intensity possibly obtainable,  $I_{TL}$  achieved by so-called *Transform-Limited* pulses, yielding  $r_t(s_t, a_t, s_{t+1}) = \frac{I_t}{I_{TL}} \in [0, 1] \ \forall t$ . In the absence of non-linear effects due to amplification, one would impose a phase on the stretcher that is opposite to the compressor's,  $\varphi_s(\omega) = -\varphi_c(\omega)$  so as to maximize intensity. As non-linearity is induced, it is reasonable to look for solutions in a neighborhood of the compressor's dispersion coefficients. Thus, one can use a multivariate normal distribution  $\mathcal{N}(-\psi_c, \epsilon \mathbb{I})$  with mean  $-\psi_c$  and diagonal variance-covariance matrix. Lastly, we employed an episodic framework for this problem, fixing the number of total interactions to T = 20, and used a discount factor of  $\gamma = 0.9$ .

#### 274 3.2 Soft Actor Critic (SAC)

Because we run training in simulation, we are able to drastically scale the experience available to the agent. With that being said, our simulation routine requires non-trivial computation, such as obtaining  $\Phi_t$  from  $\psi_t$ . Thus, we limit ourselves to the generally more sample-efficient end of DRL, and refrain from using purely on-policy methods, such as Schulman et al. (2015; 2017).

279 SAC is an off-policy DRL algorithm that leverages the power of deep function approximators to 280 learn O-functions (policy evaluation) that generalize across high-dimensional state-action spaces. 281 Then, a stochastic policy is iteratively learned by explicitly maximizing the current Q-function es-282 timate (policy improvement). Interestingly, the Q-function itself is learned in a maximum entropy 283 framework, leading to improved exploration and overall more effective learning over competing 284 methods such as DDPG (Haarnoja et al., 2018). In this work, we implement both vanilla-SAC and 285 asymmetric-SAC. The latter makes use of additional privileged information about the dynamics  $\xi$ 286 while training. Notably, this information is yet not accessible by the policy, which is only con-287 ditioned on the current state. The adoption of this asymmetric paradigm has proven empirically 288 effective in easing the training process, by providing full information to the critic networks which 289 are nevertheless not queried at test time (Akkaya et al., 2019).

#### 290 3.3 Domain Randomization (DR)

To improve on the generalization of the control policy over unknown test conditions  $\xi \sim \Xi^{real}$ , we train a control policy in simulation by sampling dynamics parameters from an arbitrary auxiliary distribution  $\Xi$ . Particularly, we compare two popular methods for choosing said distribution over  $\xi$ ,



Figure 4: SAC, learning to shape temporal pulses directly from FROG traces. The temporal profile associated with the FROG trace is superimposed on the top right of the trace for visualization purposes, and is never made available to the agent. In under 20 interactions, the agent produces near-TL temporal profiles.



Figure 5: Evolution of the controls applied by BO vs RL. As it samples from an iteratively-refined surrogate model of the unknown function  $f(\psi) = I^*$ , BO explores much more erratically than RL.

namely Uniform Domain Randomization (UDR) (Tobin et al., 2017; Sadeghi & Levine, 2016) and
 Domain Randomization via Entropy Maximization (DORAEMON) (Tiboni et al., 2023c).

UDR models  $\Xi$  as a uniform distribution over manually defined bounds  $[\xi_{\min}, \xi_{\max}]$ . Crucially, 296 297 identifying the bounds to use in training is an inherently brittle process: too-narrow bounds could 298 hinder generalization, by not providing sufficient diversity over training. On the other hand, too-299 wide bounds can yield over-regularization, and thus result in reduced performance at test time. In 300 the context of our application, experimentalists at ease with the specific pump-chain laser considered 301 in this work estimate  $B \approx B_{\text{est}} = 2$ . Thus, we train a UDR policy in simulation by using  $\xi = B \sim$ 302  $\mathcal{U}(1.5, 2.5)$ , which is roughly equivalent to allowing misspecification of up to 25% error. However, 303 even assuming access to ground-truth bounds, the probability mass of B is unlikely to be uniformly 304 distributed on large supports-this would severely impact the performance of the system on a day 305 to day basis. Conversely, it is reasonable to expect mass to be concentrated around some value 306 within a possibly larger support, further away from  $B_{est}$ . In DORAEMON Tiboni et al. (2023c), the 307 authors resolve the ambiguities in defining the training distribution by employing the principle of 308 maximum entropy (Jaynes, 1957). In other words, one could simply define a success indicator for 309 the task, and seek for the maximum entropy training distribution  $\Xi$  that satisfies a lower bound on 310 the success rate. More precisely, DORAEMON solves this problem with a curriculum of evolving 311 Beta distributions  $\Xi_k \sim \text{Beta}(a_k, b_k)$ . In line with Tiboni et al. (2023c), we apply DORAEMON as an implicit meta-learning strategy for training adaptive policies over hidden dynamics parameters. 312 We define a custom success indicator function on trajectories  $\tau_{\xi_k}$ : terminal-state pulses  $\chi(\psi_T)$  must 313 314 convey at least 65% of the TL-intensity for the respective episode to be considered successful. As a 315 result, our implementation yields an automatic curriculum over DR distributions  $\Xi$  at training time 316 such that entropy grows so long as the success rate is above 50%—as in the original paper.

#### 317 4 Experiments

318 We validate our claims on the improved machine-safety of RL over popular baselines such as 319 BO (Shalloo et al., 2020) by comparing the evolution of the controls applied at test time for the 320 both BO and mini-SAC. As BO cannot be used to process images, we benchmark it against a sim-321 plified version of our algorithm that uses exclusively  $\psi$  in the state vector, which we refer to as 322 mini-SAC. Figure 5 displays the evolution of the controls applied over the first 20 interactions be-323 tween BO and the RL-based controller. Unlike BO's solutions, which are stationary and can only be 324 transferred assuming high-fidelity simulations, RL policies can be transferred across domains and 325 adapt at test time, leveraging a sequential decision-making framework. Notably, this allows us to





Figure 6: Comparison of different strategies, measured by the average max peak intensity over 5 test episodes as a function of B-integral. These results illustrate DORAEMON's comparable performance with hand-tuned bounds for UDR.

Figure 7: Evolution of the distribution used when training an agent with DR via Entropy Maximization (DORAEMON). Later updates  $20 \ge k \ge 4$  do not further impact the evolution of the distribution over *B*.

# allocate dangerous erratic exploration to in-simulation training, severely limiting erratic exploration at test time—similarly to established work in robotics (Kober et al., 2013).

328 Since temporal profiles  $\chi(\psi)$  are typically unavailable, we exclusively use 64x64 single-channel 329 images as state representations for the agent, as discussed in 3.1. Table 1 shows the average max 330 peak intensity over 10 test episodes, after training SAC for 200k timesteps in simulation on a fixed 331  $\xi \sim \delta(B_{\text{est}})$ , while Figure 4 shows the FROG traces corresponding to the controls applied during a test episode at various timesteps. Effectively, the policy exhibits the capability of controlling  $\psi$  to 332 333 compress the pulse in time, achieving an average of 86.2% of TL's peak intensity, with peaks close 334 to 90%(2). These findings also attest the effectiveness of using single-channel images as affordable 335 proxy input to maximize peak intensity.

336 Later, we benchmark the robustness of our policy to changes in the dynamics. Particularly, we employ DR during training, and use Asymmetric-SAC together with a stack of the last n = 5 states, 337 338 vielding a history-based policy. This has shown to be effective in the context of DR to promote 339 adaptive, meta-learning behavior (Chen et al., 2021; Tiboni et al., 2023c; Akkaya et al., 2019). 340 We evaluate the performance of our method by measuring the average max intensity versus equally-341 spaced changes in the value of B-integral (cf. Figure 6). We then zoom in on these values, and report 342 in Table 1 the average peak intensity for the test conditions within [1, 3.5] (i.e., in distribution con-343 texts). When trained with DR, Asymmetric-SAC expectedly exhibits stronger robustness to changes 344 in the parametrization of the test environment. However, performance varies significantly based on 345 the distribution used while training, motivating the use for automated DR methods—Table 1 shows 346 the impact of choosing narrower rather than wider bounds for UDR, as we find wider UDR to cause 347 over-regularization, hindering performance at test time. We therefore compare the naive UDR approach with DORAEMON, by adapting the training distribution  $\{\Xi_k\}_{k=1}^K$  across K = 20 steps over 348 200k timesteps. Compared to UDR, DORAEMON displays better test-time performance around 349 our estimate  $B_{est} = 2$ , and generally provides superior success rate (cf. Table 3). Figure 7 shows 350 the evolution of the distributions  $\{\text{Beta}(a_k, b_k)\}_{k=1}^{K}$  over the course of training. Interestingly, the 351 distributions eventually converge to the maximum entropy  $\mathcal{U}(1, 3.5)$ , indicating that sufficient train-352 353 ing performance can be maintained even in the extreme case. To investigate the effectiveness of the 354 curriculum for DORAEMON, we then evaluate it against naive UDR on a slighly narrower support 355  $\mathcal{U}(1,3)$ , and observe DORAEMON's superior in Table 1).

Table 1: Average (plus-minus standard deviation) maximal peak intensity over 10 test episodes, for a combination of algorithms, training and testing conditions.  $\delta$  refers to Dirac mass, i.e. no randomization. We test our algorithms on fixed values of B.

Algorithm	Training timesteps	Training Distribution	Avg. Max Peak Intensity $(B = 1.68)$	Avg. Max Peak Intensity $(B = 2.08)$	Avg. Max Peak Intensity $(B = 2.87)$
SAC	200k	$\delta(2)$	86.18 ± 1.60	$83.80 \pm 2.34$	77.67 ± 2.53
SAC	200k	U(1.5, 2.5)	82.43 ± 5.36	$80.42 \pm 2.80$	$77.14 \pm 2.86$
SAC	200k	U(1, 3)	85.82 ± 1.48	84.85 ± 1.50	$77.71 \pm 2.18$
Asymmetric-SAC	200k	U(1.5, 2.5)	$88.69 \pm 0.60$	$86.07 \pm 0.49$	$79.32 \pm 1.12$
Asymmetric-SAC	200k	DORAEMON(1, 3.5)	$86.04 \pm 3.78$	$85.12 \pm 1.10$	$79.34 \pm 1.59$

Table 2: Min-Max ranges for the maximal peak intensity over 10 test episodes, for a combination of algorithms, training and testing conditions.

Algorithm	Train timesteps	Train Distribution	<b>Min-Max</b> <b>Peak Intensity</b> $(B = 1.68)$	<b>Min-Max</b> <b>Peak Intensity</b> $(B = 2.08)$	<b>Min-Max</b> <b>Peak Intensity</b> $(B = 2.87)$
SAC	200k	$\delta(2)$	83.95-89.13	79.87-86.38	72.65-80.69
SAC	200k	U(1.5, 2.5)	69.04-89.23	74.99-84.07	71.16-80.35
SAC	200k	$\mathcal{U}(1,3)$	83.35-87.65	82.07-86.19	74.87-80.03
Asymmetric-SAC	200k	U(1.5, 2.5)	87.26-89.31	84.76-86.39	77.15-80.53
Asymmetric-SAC	200k	DORAEMON(1, 3.5)	76.24-89.37	83.17-86.27	75.04-80.77

Table 3: Success rate over 10 test episodes: proportion of episodes with a maximal peak intensity  $\geq 80\%$  of TL in multiple experimental conditions. DORAEMON shows to be best suited to tackle more challenging scenarios with more pronounced non-linear effects compared to UDR.

Train Distribution	Success Rate	Success Rate	Success Rate
IT all Distribution	(B = 1.68)	(B = 2.08)	(B = 2.87)
$\delta(2)$	1.0	0.9	0.2
U(1.5, 2.5)	0.9	0.6	0.1
$\mathcal{U}(1,3)$	0.5	0.5	0.1
U(1.5, 2.5)	1.0	1.0	0.2
DORAEMON(1, 3.5)	0.9	1.0	0.4
	$\begin{array}{c} \delta(2) \\ \mathcal{U}(1.5,2.5) \\ \mathcal{U}(1.3) \\ \mathcal{U}(1.5,2.5) \\ \mathcal{U}(1.5,2.5) \\ \text{DORAEMON}(1,3.5) \end{array}$	$\begin{tabular}{lllllllllllllllllllllllllllllllllll$	$\begin{tabular}{ c c c c } \hline {\bf Train Distribution} & Success Rate \\ (B = 1.68) & (B = 2.08) \\ \hline $\delta(2)$ & 1.0$ & 0.9 \\ $U(1.5, 2.5)$ & 0.9$ & 0.6 \\ $U(1.5, 2.5)$ & 1.0$ & 0.5 \\ $U(1.5, 2.5)$ & 1.0$ & 1.0 \\ \hline $DORAEMON(1, 3.5)$ & 0.9$ & 1.0 \\ \hline \end{tabular}$

## 356 **5** Conclusions

In this work, we present a novel application of RL to the rich and complex domain of experimental laser physics, using RL as the backbone for a fully automated pulse-shaping routine. Leveraging domain knowledge of the processes regulating phase accumulation in HPL systems, we design a coarse simulator of the pump chain of a HPL system, and we use it to develop control strategies that exclusively use non-destructive measurements in the form of images to maximize the peak intensity of ultra-short laser pulses.

363 We benchmark our method against popular black-box approaches to pulse intensity maximization 364 (i.e. duration minimization), and argue that our approach is inherently better suited for real-world applications as it can learn to apply gentle controls not endangering system safety, and produce peak 365 366 intensities as high as 90% of TL's. Further, we reformulate the problem of pulse shaping as a Latent 367 MDP, and employ the latest advancements in the field of Domain Randomization to develop adaptive 368 policies capable of producing ultra-short laser pulses for a wide range of dynamics parameters. Our 369 work is a concrete step towards the application of DRL to controlling HPL systems, with the goal 370 of streamlining the production of and advancing studies on ultra-short laser pulses and extreme light-matter interactions. 371

**Limitations.** We identify several limitations remaining in our contribution. In particular, HPL systems' performance is known to be influenced, alongside B-integral, by the dispersion coefficients of the compressor. These dispersion coefficients are highly sensitive to the delicate alignment of the compressor optics, which is typically a cumbersome and time-consuming process in ultra-fast optics. As such, we concluded randomizing over these coefficients was unnecessary in a first instance, as a great deal of effort and diagnostic is spent in properly assessing and monitoring the compressor. Still, adapting to their variation as well is a very promising approach, which we seek to investigate 379 further.

380 Another limitation is the sample inefficiency of our method, requiring hundreds of thousands to sam-

381 ples to discover well performing policies. We argue this is particularly problematic considering the

382 knowledge available on the process of phase accumulation in linear and non-linear crystals. While

383 our coarse simulator provides a useful tool for model-free learning, the absence of explicit model-

ing of the dynamics limits data efficiency. Integrating model-based components could significantly improve sample efficiency.

386 Despite these limitations, our work takes a significant step toward the integration of DRL in HPL

- 387 systems, providing a framework that is both practical and adaptable to experimental constraints, and
- 388 prove the effectiveness of the technique in ultra-short laser physics.

# 389 **References**

H Abu-Shawareb, R Acree, P Adams, J Adams, B Addis, R Aden, P Adrian, BB Afeyan, M Aggle ton, L Aghaian, et al. Achievement of target gain larger than unity in an inertial fusion experiment.
 *Physical review letters*, 132(6):065102, 2024.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron,
 Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a
 robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Ishraq Md Anjum, Davorin Peceli, Francesco Capuano, and Bedrich Rus. High-power laser pulse
 shape optimization with hybrid stochastic optimization algorithms. In *Laser Science*, pp. JD4A–
 55. Optica Publishing Group, 2024.

Rika Antonova, Silvia Cruciani, Christian Smith, and Danica Kragic. Reinforcement learning for
pivoting task. *arXiv preprint arXiv:1703.00472*, 2017.

Francisco Rodrigo Arteaga-Sierra, C Milián, I Torres-Gómez, M Torres-Cisneros, Germán Moltó,
and A Ferrando. Supercontinuum optimization for dual-soliton based light sources using genetic
algorithms in a grid platform. *Optics express*, 22(19):23686–23693, 2014.

T Baumert, T Brixner, V Seyfried, M Strehle, and G Gerber. Femtosecond pulse shaping by an evolutionary algorithm with feedback. *Applied Physics B: Lasers & Optics*, 65(6), 1997.

Francesco Capuano, Davorin Peceli, Gabriele Tiboni, Alexandr Špaček, and Bedřic Rus. Laser pulse
duration optimization with numerical methods. In *Proceedings of the PCaPAC2022 conference*,
pp. 37–40. JaCoW, 2022.

- Francesco Capuano, Davorin Peceli, Gabriele Tiboni, Raffaello Camoriano, and Bedřich Rus. Temporl: laser pulse temporal shape optimization with deep reinforcement learning. In *High-power*, *High-energy Lasers and Ultrafast Optical Technologies*, volume 12577, pp. 62–74. SPIE, 2023.
- Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding domain randomization for sim-to-real transfer. *arXiv preprint arXiv:2110.03239*, 2021.

Thomas Gaumnitz, Arohi Jain, Yoann Pertot, Martin Huppert, Inga Jordan, Fernando ArdanaLamas, and Hans Jakob Wörner. Streaking of 43-attosecond soft-x-ray pulses generated by a
passively cep-stable mid-infrared driver. *Optics express*, 25(22):27506–27518, 2017.

- Gabriele Maria Grittani, Tadzio Levato, Carlo Maria Lazzarini, and Georg Korn. Device and
  method for high dose per pulse radiotherapy with real time imaging, March 31 2020. US Patent
  10,603,514.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy
   maximum entropy deep reinforcement learning with a stochastic actor. In *International confer- ence on machine learning*, pp. 1861–1870. Pmlr, 2018.

- 423 Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
- Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.
- Evgeny Kuprikov, Alexey Kokhanovskiy, Kirill Serebrennikov, and Sergey Turitsyn. Deep reinforcement learning for self-tuning laser source of dissipative solitons. *Scientific Reports*, 12(1):
  7185, 2022.
- B Loughran, MJV Streeter, H Ahmed, S Astbury, M Balcazar, M Borghesi, N Bourgeois, CB Curry,
  SJD Dann, S Dilorio, et al. Automated control and optimisation of laser driven ion acceleration. *High Power Laser Science and Engineering*, pp. 1–11, 2023.
- Evgenii Mareev, Alena Garmatina, Timur Semenov, Nika Asharchuk, Vladimir Rovenko, and Irina
  Dyachkova. Self-adjusting optical systems based on reinforcement learning. In *Photonics*, volume 10, pp. 1097. MDPI, 2023.
- Gabriel B Margolis, Ge Yang, Kartik Paigwar, Tao Chen, and Pulkit Agrawal. Rapid locomotion via
  reinforcement learning. *The International Journal of Robotics Research*, 43(4):572–587, 2024.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- 440 Rüdiger Paschotta. *Field guide to laser pulse generation*, volume 14. SPIE press Bellingham, 2008.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of
robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

- Ildar Rakhmatulin, Donald Risbridger, RM Carter, MJ Daniel Esser, and Mustafa Suphi Erden.
  Reinforcement learning for aligning laser optics with kinematic mounts. In 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE), pp. 1397–1402. IEEE, 2024.
- Fereshteh Sadeghi and Sergey Levine. Cad2rl: Real single-image flight without a single real image.
   *arXiv preprint arXiv:1611.04201*, 2016.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region
  policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR,
  2015.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- RJ Shalloo, SJD Dann, J-N Gruse, CID Underwood, AF Antoine, Christopher Arran, Michael Backhouse, CD Baird, MD Balcazar, Nicholas Bourgeois, et al. Automation and control of laser
  wakefield accelerators using bayesian optimization. *Nature communications*, 11(1):6355, 2020.
- Gabriele Tiboni, Karol Arndt, Giuseppe Averta, Ville Kyrki, and Tatiana Tommasi. Online vs.
  offline adaptive domain randomization benchmark. In Pablo Borja, Cosimo Della Santina, Luka
  Peternel, and Elena Torta (eds.), *Human-Friendly Robotics 2022*, pp. 158–173, Cham, 2023a.
  Springer International Publishing. ISBN 978-3-031-22731-8.
- Gabriele Tiboni, Karol Arndt, and Ville Kyrki. Dropo: Sim-to-real transfer with offline domain
  randomization. *Robotics and Autonomous Systems*, pp. 104432, 2023b. ISSN 0921-8890. DOI: https://doi.org/10.1016/j.robot.2023.104432.

Gabriele Tiboni, Pascal Klink, Jan Peters, Tatiana Tommasi, Carlo D'Eramo, and Georgia Chalvatzaki. Domain randomization via entropy maximization. *arXiv preprint arXiv:2311.01885*, 2023c.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In
2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp. 23–30.
IEEE, 2017.

472 Rick Trebino and Daniel J Kane. Using phase retrieval to measure the intensity and phase of ultra473 short pulses: frequency-resolved optical gating. *Journal of the Optical society of America A*, 10
474 (5):1101–1111, 1993.

475 Rick Trebino, Kenneth W DeLong, David N Fittinghoff, John N Sweetser, Marco A Krumbügel,
476 Bruce A Richman, and Daniel J Kane. Measuring ultrashort laser pulses in the time-frequency
477 domain using frequency-resolved optical gating. *Review of Scientific Instruments*, 68(9):3277–
478 3295, 1997.

Eugene Valassakis, Zihan Ding, and Edward Johns. Crossing the gap: A deep dive into zero-shot
sim-to-real transfer for dynamics. In 2020 IEEE/RSJ International Conference on Intelligent *Robots and Systems (IROS)*, pp. 5372–5379. IEEE, 2020.

RI Woodward and Edmund JR Kelleher. Towards 'smart lasers': self-optimisation of an ultrafast
pulse source using a genetic algorithm. *Scientific reports*, 6(1):1–9, 2016.

Tom Zahavy, Alex Dikopoltsev, Daniel Moss, Gil Ilan Haham, Oren Cohen, Shie Mannor, and Mordechai Segev. Deep learning reconstruction of ultrashort pulses. *Optica*, 5(5):666–673, 2018.

Shaojun Zhu, Andrew Kimmel, Kostas E Bekris, and Abdeslam Boularias. Fast model identification
via physics engines for data-efficient policy search. *arXiv preprint arXiv:1710.08893*, 2017.