

Scaling Scaling Laws with Board Games

Andy L. Jones

London, United Kingdom
me@andyjones.com

Abstract—The largest experiments in machine learning now require resources far beyond the budget of all but a few institutions. Fortunately, it has recently been shown that the results of these huge experiments can often be extrapolated from the results of a sequence of far smaller, cheaper experiments. In this work, we show that not only can the extrapolation be done based on the size of the model, but on the size of the problem as well. By conducting a sequence of experiments using AlphaZero and Hex, we show that the performance achievable with a fixed amount of compute degrades predictably as the game gets larger and harder. Along with our main result, we further show that the test-time and train-time compute available to an agent can be traded off while maintaining performance.

Index Terms—Scaling Laws, Deep Reinforcement Learning

I. INTRODUCTION

There is a concern that the state-of-the-art models studied by the most well-resourced organisations are growing too expensive for other researchers to keep pace [1]–[3]. Fortunately, the recently-proposed paradigm of *scaling laws* proposes a solution: that by studying the behaviour of a sequence of small, cheap models, researchers can extrapolate the behaviour of large, expensive models without having to explicitly train them.

In the past year, scaling laws have been established over a range of domains in machine learning [4]–[9]. These laws show that the performance of each model in a family can be well-characterised by a function some ‘size’ property (like data or compute), and that the function behaves predictably over many orders of magnitude in model size.

So far however these works have only considered scaling the size of the model, leaving fixed the problem under consideration. Our principal contribution is to generalise this, scaling not only the model but the problem as well. In this work, we show that the behaviour of a model on a small problem instance predicts the behaviour of a model on a much larger problem instance.

Our problem of choice is the board game Hex [10], a strategic board game whose complexity can be easily adjusted by changing the board size. Using AlphaZero [11], we train many different models on many different board sizes. Analysed together, the performance of these models reveals a *compute frontier* that bounds the performance a model from our family in terms of the compute used to train it. These compute frontiers are exponential in the desired performance, and exponential again in the board size.

Building on these results, we show that compute frontiers fitted at small board sizes are good predictors of the compute

frontiers discovered at large board sizes. More, the error in the prediction drops exponentially as more small board sizes are added to the fit.

Finally, while pursuing our main results we discovered an independently-interesting result: that for each extra order of magnitude of train-time compute, we can reduce test-time compute by a similar factor while leaving performance unchanged.

We have published our code, models and data on GitHub¹.

II. BACKGROUND

A. Scaling Laws

While the general idea of studying power laws in model size stretches back to at least the 1980s [12], it was the work of Hestness et al. [4] that first brought the phenomenon to the attention of a contemporary audience. Their work showed that over a range of network architectures, the performance of a language model followed a power-law in the size of the dataset it was trained on.

Later, Rosenfeld et al. [5] showed that the fit of the power law could be substantially improved by taking into account the size of the model, while Kaplan et al. [6] further added the amount of compute spent training it. Then in Henighan et al. [7], these laws were further shown to hold – with varying coefficients – over a range of generative modelling tasks, including video. Most recently Hernandez et al. [9] have shown laws in fine-tuning, and Rosenfeld et al. [8] in pruning.

There has also been work on the theoretical underpinnings of these laws. Hutter [13] is the most recent contribution in the area, and its introduction provides an exhaustive overview of prior work.

So far however, published work on scaling laws has exclusively addressed images and language. The forthcoming Hilton et al. [14] studies scaling laws in single-agent reinforcement learning, but ours is the first work on scaling laws in multi-agent reinforcement learning, and the first to scale the size of the problem as well as the size of the model.

B. AlphaZero

AlphaZero [11] is an algorithm for teaching a neural network to play a two-player zero-sum game entirely through self-play. At each step in the training process, AlphaZero augments the network-generated policy with a tree search. The augmented policy is stronger than the original policy

¹<https://andyjones.com/boardlaw/>

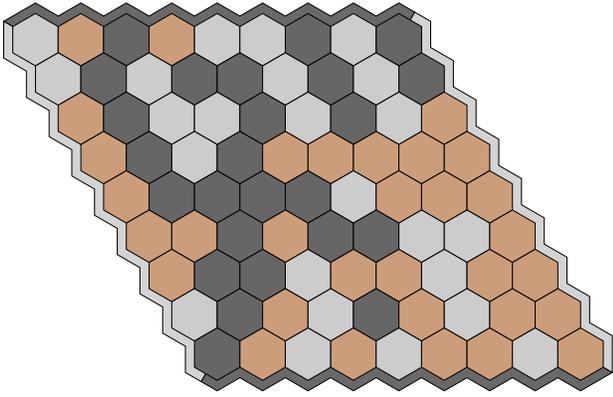


Fig. 1. A Hex game on a 9×9 board, won by black with the path in the second column.

on its own, and consequently self-play games between the augmented network and itself can be used as a source of experience to train the network. This amplification process [15] progressively bootstraps the network from a random initialisation up to superhuman play, and - importantly - does so in a way that requires no extra human input.

C. Hex

Hex [10] is a strategy game for two players. The players take turns placing tokens on a rhombic board, and the first player to connect their sides of the board is the winner (Fig 1). First developed by Hein in 1942 [16], Hex has enjoyed niche popularity throughout its life [17].

Despite the simplicity of its rule set, Hex is considered to be a complex strategic game [18]. In fact, despite sustained attention from games researchers [19]–[21], computers only surpassed human-level play at large board sizes in 2020 [22].

We chose Hex as the focus of our work because it is easy to vary the size and complexity of the game, and because it is easy to implement as a fast, parallelized GPU kernel. More popular games such as Chess, Go and Shogi have all accumulated minor rules - such as castling, *kō* or *nifu* - that make for dramatically more complex and bug-prone implementations [23].

One further simplification we make is that while human games of Hex are typically played with the ‘pie rule’ as a way to nullify first-mover advantage, in our implementation we omit it. Instead, all evaluation matches are played as a pair, with each agent playing black in one match and white in the other.

D. Ratings and Elo

Unlike in regular reinforcement learning where performance (reward) can be measured easily, the performance of an agent in a multiplayer game depends on who the opponent is. As such, any rating system needs to take account of not only the player but also their opponent.

In human chess tournaments, the solution is the Elo system [24]. The Elo system assigns each player a numerical ranking - their *Elo* - in such a way that that the chance of one player

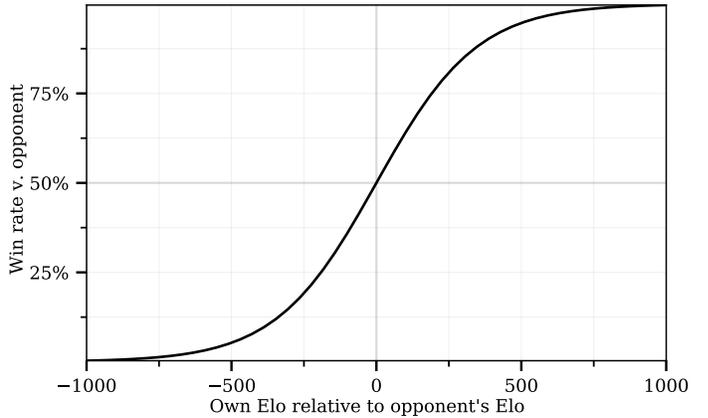


Fig. 2. The Elo ratings of two players predict the outcome of a match between them, with the player with the higher Elo being more likely to win.

winning over another can be calculated from the difference between the two player’s Elos (Fig 2). Stronger players come out of this system with high Elos; weak players with low Elos.

The central limitation of the Elo system is that it assumes *transitivity*. This is not necessarily the case, and in fact there are games - such as rock-paper-scissors - where the Elos assigned to each player are entirely uninformative [25]–[27].

Elo is also a *relative* rating system, meaning that any set of Elo ratings can be shifted up or down by a constant offset without affecting their predictive ability. Fortunately, on our board sizes there is an excellent choice of constant offset: fixing perfect play to zero Elo. MoHex [19], [28]–[30] is an algorithmic agent that can play perfectly on board sizes up to 9×9 , and we fix its play to zero for all Elo ratings reported herein.

While Elo is the best known rating system of its type, there are other more modern variations such as Glicko [31] and TrueSkill [32]. These variations are all more complex however, and the additional complexities would not improve the analyses carried out in this work.

III. METHODS

We developed a fast, low-resource AlphaZero implementation (documented in Appendix A) and used it to train many different models on many different board sizes. We then evaluated the trained models against perfect play in order to come up with compute frontiers at each board size. Finally, we fitted a simple curve to these frontiers, to show that the relationship is consistent across board sizes.

A. AlphaZero

Our implementation of AlphaZero can train an agent to perfect play in time approximately exponential in board size (Fig 3). In particular, perfect play on a 9×9 board takes a little under 3 hours on a single RTX 2080 Ti. We have not been able to find good baselines for the performance of our implementation - the only other 9×9 AlphaZero Hex implementation we know of is Polygames’ [22], and training time figures have not been made available for it.

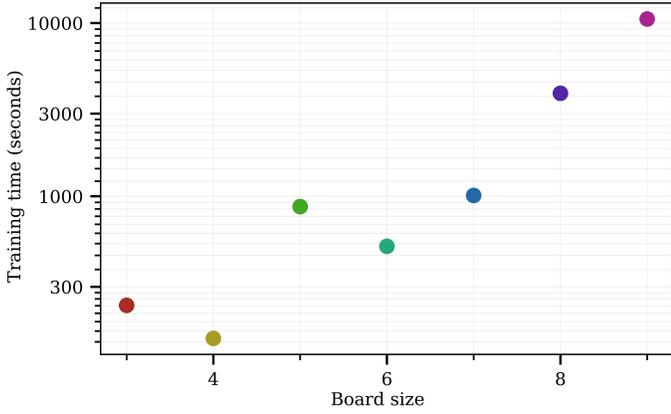


Fig. 3. The time taken to train an agent to -50 Elo (ie, almost equal to perfect play) is roughly exponential in boardsize, with the fastest agent on a 9×9 board taking about 3 hours.

TABLE I
HYPERPARAMETERS

Number of envs	32k
Batch size	32k
Buffer size	2m samples
Learning rate	$1e-3$
MCTS node count	64
MCTS c_{puct}	$1/16$
MCTS noise ϵ	$1/4$

B. Models

We used AlphaZero to train ≈ 200 different models over a range of hyperparameters. Most hyperparameters were held constant across runs and are documented in Table I, while a few - principally the network architectures and run duration - varied with the board size, and are documented in Table II.

The independent variables for our analysis are board size and compute. Varying board size is simple, but there are many ways to vary the amount of compute involved in training a model. We chose to explore three axes of compute variation: the depth of the network, the width of the network, and the length of the training run. Specifically,

1) *Board size*: Board sizes ranged from 3 to 9. The smallest board used was 3×3 , as this is the smallest ‘interesting’ board size. The largest board used was 9×9 , as this was the largest board MoHex can achieve perfect play on.

TABLE II
BOARD SIZE-DEPENDENT HYPERPARAMETER LIMITS

Board Size	Neurons	Layers	Samples	Compute
3	2	4	$4E+08$	$1E+12$
4	16	4	$2E+08$	$1E+13$
5	16	8	$3E+08$	$3E+13$
6	128	8	$6E+08$	$4E+14$
7	512	8	$1E+09$	$1E+16$
8	512	8	$1E+09$	$3E+16$
9	1024	8	$2E+09$	$1E+17$

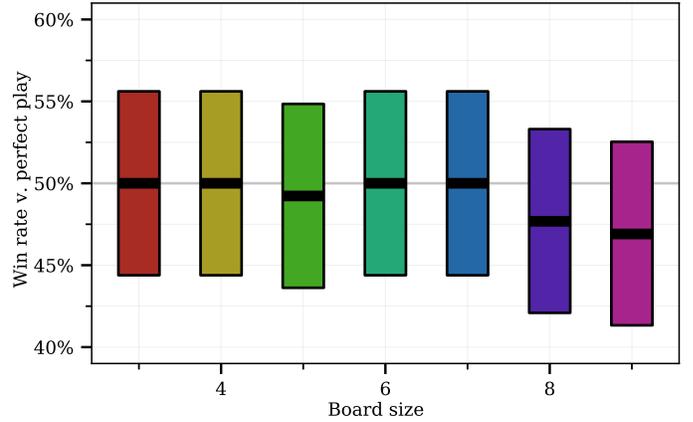


Fig. 4. Our best AlphaZero agents are on par with MoHex’s perfect play. Shown are the 90% credible intervals on the best agents’ win rate against MoHex after 128 games, assuming a Beta(1, 1) prior.

2) *Agent architecture*: Agent architectures ranged in powers of 2 from 1 layer of 1 neuron through to 8 layers of 1024 neurons. The maximum agent size for each board size was determined during preliminary work, and is listed in Table II.

3) *Run length*: Training runs were terminated when they hit a certain number of samples or a certain number of FLOPS-seconds. These limits were also determined during preliminary work, and are listed in Table II.

4) *Snapshots*: Snapshots were taken from the training run on a schedule exponential in compute. The schedule was chosen so that a training run hitting the compute limit would have 21 snapshots taken. Overall, we took 2,800 snapshots in total.

C. Evaluation

We evaluated the agents by playing each agent against each other agent for 1024 matches, with each agent playing black for 512 of those matches and white for the other 512. We chose this number of matches based on hardware, time constraints, and the number of pairings that needed to be evaluated. We then used the outcomes from the matches to calculate an Elo rating for each agent.

Playing 1,024 matches between each pair of snapshots means playing 700m matches overall. To accelerate the evaluation, we took groups of 64 agents and played all 2m matches between them in parallel, batching the inferences for evaluation on the GPU. By fully saturating the GPU, we found we could play about 1k evaluation matches/GPU/second.

While the matches between AlphaZero agents can establish the relative ratings, to fix the offset we also played the top-ranking agents against MoHex. The top-ranking agents reliably draw MoHex (Fig. 4), showing they are on par with perfect play.

1) *Hyperparameters*: The same search hyperparameters were used for evaluation as were used during training, as listed in Table I.

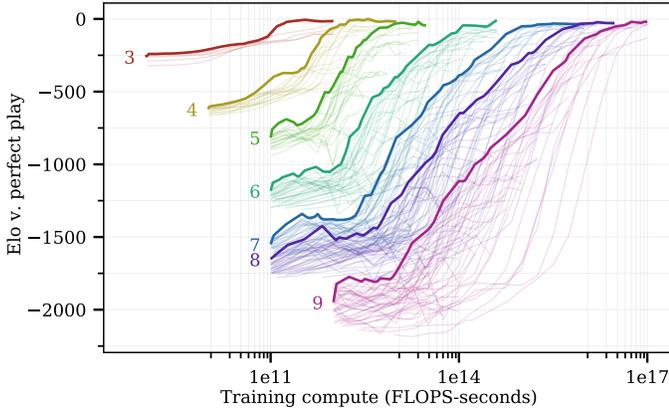


Fig. 5. Each training run (each faint line) of each differently-sized agent follows a sigmoid, starting at random play and progressing up to some plateau. The frontiers (dark lines) formed by taking a maximum across training runs have a similar form across board sizes (colors).

D. Hardware

Each training run was conducted on a single RTX 2080 Ti, with many runs being carried out in parallel on machines rented from vast.ai. In all, about 500 GPU-hours were used for training.

Evaluation matches meanwhile were carried out on two in-house RTX 2080 Tis, taking about 100 GPU-hours in all.

E. Curve fitting

Having trained and evaluated the agents, the final step is to fit a functional form to the frontiers. The frontiers give the maximum performance attained for each quantity of compute at each board size, and can be roughly described as sequence of parallel plateaus, leading up into a set of parallel inclines, leading out onto a second plateau at zero Elo.

We explored several formalisations of this pattern (Appendix C) before settling on a five-parameter change-point model:

$$\begin{aligned} \text{plateau} &= m_{\text{boardsize}}^{\text{plateau}} \cdot \text{boardsize} + c^{\text{plateau}} \\ \text{incline} &= m_{\text{boardsize}}^{\text{incline}} \cdot \text{boardsize} + m_{\text{flops}}^{\text{incline}} \cdot \log \text{flop} + c^{\text{incline}} \\ \text{elo} &= \text{incline}.\text{clamp}(\text{plateau}, 0) \end{aligned}$$

The first equation gives the lower set of parallel plateaus, the second the parallel inclines, and the third combines them. We fit the model with L-BFGS.

IV. RESULTS

A. Frontier parameters

During training, the performance of each agent describes a rough sigmoid in terms of compute spent (Fig. 5). Taking the maximum across agents at each level of compute gives the compute frontier, to which we fit our change-point model.

The fitted frontiers are shown in Fig. 6, and the parameters of those fits in Table III. These parameters are easier to understand in terms of derived quantities:

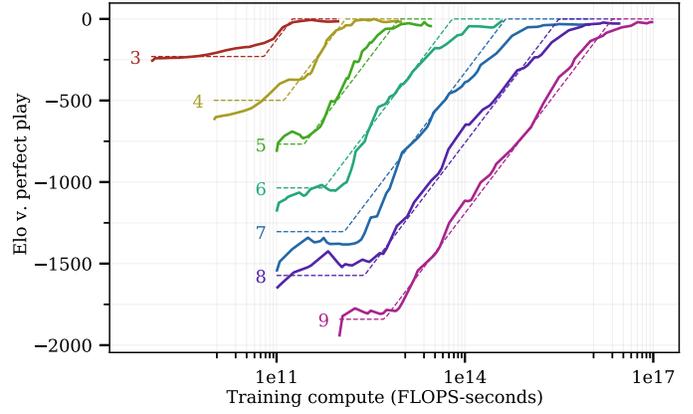


Fig. 6. The compute-performance frontier follows the same sigmoid for each board size 3 through 9, just scaled and shifted. The dotted lines give the fitted curves.

TABLE III
FITTED FRONTIER PARAMETERS

	m_{flops}	$m_{\text{boardsize}}$	c
plateau		-270	570
incline	510	-430	-4400

1) *Slope*: The slope of the incline is 500 Elo per order of magnitude increase in compute. A more memorable interpretation is that if you are in the linearly-increasing regime, then you will need about $2\times$ as much compute as your opponent to beat them $2/3$ of the time.

2) *Perfect play*: The minimum compute needed for perfect play increases $7\times$ for each increment in board size.

3) *Takeoff*: The minimum training compute needed to see any improvement over random play increases by $4\times$ for each increment of board size.

4) *Random play*: Finally, the distance between random play and perfect play increases by 500 Elo for each increment of board size. Unlike the other quantities mentioned previously, the distance between random and perfect play is a property of the game itself rather than of the agent.

B. Predictive errors

While the model in the previous section was fitted across all board sizes simultaneously, we can alternatively ask: if we fit the model on data up to some small board size, how well does the fit predict the data from higher, unseen board sizes?

As can be seen in Fig. 7, the frontiers found at smaller board sizes accurately predict the frontiers that will be found at larger board sizes. The error in the predicted frontier (as measured by the residual variance) starts small and decays exponentially as more small boards are added to the fit.

C. Train-test trade-off

While developing main results discussed above, a small unit of extra work was suggested towards an independently interesting result².

²Thanks and credit to Jared Kaplan for suggesting this.

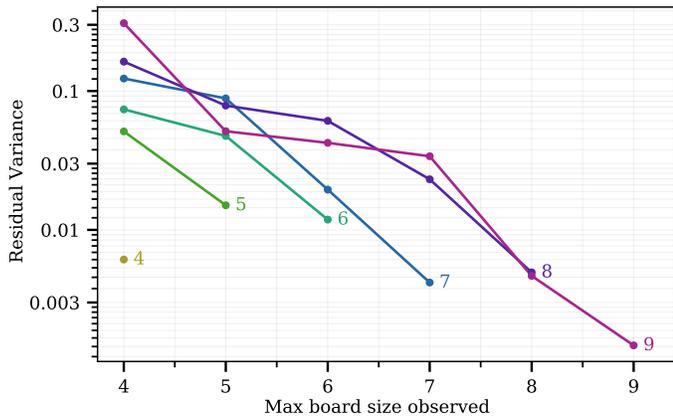


Fig. 7. The error in the prediction decays exponentially as more boards are used. Each line gives the errors in the prediction for the frontier of a specific board size.

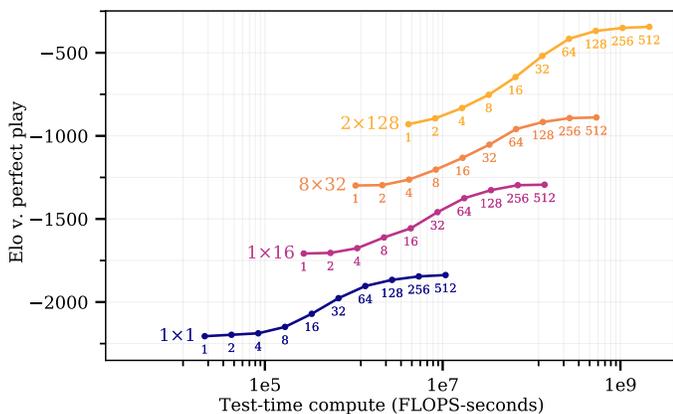


Fig. 8. A selection of snapshots trained on a 9×9 board, evaluated with varying test-time tree sizes. These curves show that the performance of a specific snapshot is sigmoid in the test-time compute budget. The lines are labelled with the architecture of the snapshot, in the format depth \times width. Each point on the line is the Elo of that snapshot evaluated with a different tree size, spaced logarithmically between 1 node and 512 nodes.

So far we have focused on the compute budget during training, but another pertinent budget is the compute spent during evaluation. All the results discussed previously have used a tree search of size 64 during evaluation, the same as used during training. But there is no reason that the train-time search and test-time search have to be the same size, and so by varying the size of the test-time compute budget we can see in Fig. 8 that larger tree searches at test time can substantially improve the performance of an agent.

Knowing now that compute can be spent in two places, at train time and test time, the immediate question is: how do these two budgets trade off? This is illustrated in Fig. 9, which shows that the trade-off is linear in log-compute: for each additional $10\times$ of train-time compute, about $15\times$ of test-time compute can be eliminated, down to a floor of a single-node tree search.

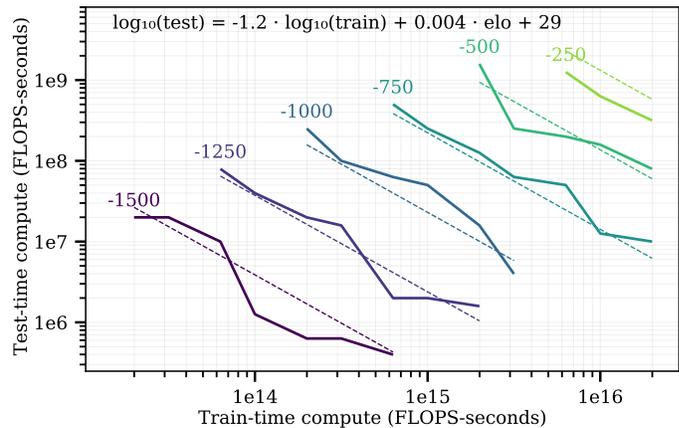


Fig. 9. The trade-off between train-time compute and test-time compute. Each dotted line gives the minimum train-test compute required for a certain Elo on a 9×9 board

V. DISCUSSION

Our central, concrete result is that when we train AlphaZero to play Hex, the compute required can be calculated directly from the board size and the desired performance. We have also shown that compute during training and compute at test time can be traded off according to simple relationship. These results illuminate several intriguing phenomena.

First, the way in which performance scales with compute is that an agent with twice as much compute as its opponent can win roughly $2/3$ of the time. This behaviour is strikingly similar to that of a toy model where each player chooses as many random numbers as they have compute, and the player with the highest number wins³. In this toy model, doubling your compute doubles how many random numbers you draw, and the probability that you possess the largest number is $2/3$. This suggests that the complex game play of Hex might actually reduce to each agent having a ‘pool’ of strategies proportional to its compute, and whoever picks the better strategy wins. While on the basis of the evidence presented herein we can only consider this to be serendipity, we are keen to see whether the same behaviour holds in other games.

Second, both the relation of performance to board size and the relation of performance to compute are smooth. Before embarking on this project, a key unknown was whether performance would show any ‘spikes’ with regards to compute or board size. A spike with regards to compute might indicate the model had achieved some key insight, while a spike with regards to board size might indicate a minimum complexity past which key insights are available for the model to discover. As is however, models’ performance changes smoothly and predictably with both increased compute and increased complexity. However, this could plausibly be a property unique to Hex and it’s simple rule set, and we would again be keen to see whether the same behaviour holds in other games.

Finally, the simple relationship between compute at train time and compute at test time was originally surprising to us.

³Thanks and credit to Paul Christiano for making us aware of this.

Our intuition was that test-time compute is much ‘cheaper’ than train-time compute, and so we were surprised that one could easily substitute for the other. On reflection however, we believe the key distinction is that an optimization at test-time needs only optimise over one sample, while train-time compute meanwhile must optimise over the entire distribution of samples.

In all, these results demonstrate how a relationship between compute and performance identified in small, cheap problems carries directly over to problems sizes that are orders of magnitude more expensive to explore. If this phenomenon proves to be general, it opens the way for researchers to contribute to the understanding of problems far larger than the ones they themselves are able to directly study.

ACKNOWLEDGEMENTS

This work was funded by Survival & Flourishing. This work has also benefited greatly from the advice of many friends and colleagues. In particular, we wish to acknowledge the invaluable input of Jared Kaplan, Jan Leike, Paul Christiano, Danny Hernandez, Jacob Hilton, Matthew Rahtz, Marc Lanctot, Max O. Smith, Ryan Hayward, Paul Lu, Adam Gleave, Asya Bergal, Mario Lezcano Casado, Ben Wang, Jeremy Salwen, Clemens Winter, and Ella Guest.

REFERENCES

[1] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” *arXiv preprint arXiv:1906.02243*, 2019.

[2] NRC Letter Signatories, “National Research Cloud Call To Action,” 2020. [Online]. Available: <https://hai.stanford.edu/national-research-cloud-joint-letter>.

[3] UK Research and Innovation, “Transforming our world with AI,” 2021.

[4] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou, “Deep learning scaling is predictable, empirically,” *arXiv preprint arXiv:1712.00409*, 2017.

[5] J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, and N. Shavit, “A constructive prediction of the generalization error across scales,” *arXiv preprint arXiv:1909.12673*, 2019.

[6] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.

[7] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, *et al.*, “Scaling laws for autoregressive generative modeling,” *arXiv preprint arXiv:2010.14701*, 2020.

[8] J. S. Rosenfeld, J. Frankle, M. Carbin, and N. Shavit, “On the predictability of pruning across scales,” *arXiv preprint arXiv:2006.10621*, 2020.

[9] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish, “Scaling laws for transfer,” *arXiv preprint arXiv:2102.01293*, 2021.

[10] Wikipedia contributors, “Hex — Wikipedia, the free encyclopedia,” 2020. [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Hex%5C&oldid=996842461>.

[11] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, “A general reinforcement learning algorithm that masters chess, shogi, and go through self-play,” *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

[12] R. A. DeVore, R. Howard, and C. Micchelli, “Optimal nonlinear approximation,” *Manuscripta mathematica*, vol. 63, no. 4, pp. 469–478, 1989.

[13] M. Hutter, “Learning curve theory,” *arXiv preprint arXiv:2102.04074*, 2021.

[14] J. Hilton and J. Tang, “Scaling laws for reinforcement learning,” In preparation.

[15] P. Christiano, “AlphaGo Zero and capability amplification,” 2019. [Online]. Available: <https://www.lesswrong.com/posts/HA3oArypzNANvXC38/alphago-zero-and-capability-amplification>.

[16] P. Hein, “Vil de laere polygon?” *Politiken*, vol. December 26, 1942.

[17] R. B. Hayward and B. Toft, *Hex: The Full Story*. CRC Press, 2019.

[18] M. Seymour, “Hex: A strategy guide,” 2020. [Online]. Available: http://www.mseymour.ca/hex_book/hexstrat.html.

[19] S.-C. Huang, B. Arneson, R. B. Hayward, M. Müller, and J. Pawlewicz, “Mohex 2.0: A pattern-based mcts hex player,” in *International Conference on Computers and Games*, Springer, 2013, pp. 60–71.

[20] K. Young, G. Vasan, and R. Hayward, “Neurohex: A deep q-learning hex agent,” in *Computer Games*, Springer, 2016, pp. 3–18.

[21] T. Anthony, Z. Tian, and D. Barber, “Thinking fast and slow with deep learning and tree search,” *arXiv preprint arXiv:1705.08439*, 2017.

[22] T. Cazenave, Y.-C. Chen, G.-W. Chen, S.-Y. Chen, X.-D. Chiu, J. Dehos, M. Elsa, Q. Gong, H. Hu, V. Khalidov, C.-L. Li, H.-I. Lin, Y.-J. Lin, X. Martinet, V. Mella, J. Rapin, B. Roziere, G. Synnaeve, F. Teytaud, O. Teytaud, S.-C. Ye, Y.-J. Ye, S.-J. Yen, and S. Zagoruyko, “Polygames: Improved zero learning,” 2020. arXiv: 2001.09832 [cs.LG].

[23] ChessProgramming contributors, “Engine testing,” 2020. [Online]. Available: https://www.chessprogramming.org/Engine%5C_Testing.

[24] A. E. Elo, *The rating of chessplayers, past and present*. Arco Pub., 1978.

[25] D. Balduzzi, K. Tuyls, J. Perolat, and T. Graepel, “Re-evaluating evaluation,” 2018. arXiv: 1806.02643 [cs.LG].

[26] M. Rowland, S. Omidshafiei, K. Tuyls, J. Perolat, M. Valko, G. Piliouras, and R. Munos, “Multiagent evaluation under incomplete information,” 2020. arXiv: 1909.09849 [cs.MA].

[27] W. M. Czarnecki, G. Gidel, B. Tracey, K. Tuyls, S. Omidshafiei, D. Balduzzi, and M. Jaderberg, “Real world games look like spinning tops,” 2020. arXiv: 2004.09468 [cs.LG].

[28] Henderson, Philip and Arneson, Broderick and Pawlewicz, Jakob and Huang, Aja and Young, Kenny and Gao, Chao, “Mohex,” version d450c01, Feb. 25, 2020. [Online]. Available: <https://github.com/cgao3/benzene-vanilla-cmake>.

[29] J. Pawlewicz, R. Hayward, P. Henderson, and B. Arneson, “Stronger virtual connections in hex,” *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 7, no. 2, pp. 156–166, 2014.

[30] J. Pawlewicz and R. B. Hayward, “Scalable parallel dfpn search,” in *International Conference on Computers and Games*, Springer, 2013, pp. 138–150.

[31] M. E. Glickman, “The glicko system,” *Boston University*, vol. 16, pp. 16–17, 1995.

[32] T. Minka, T. Graepel, and R. Herbrich, “Trueskill tm: A bayesian skill rating system,” *Advances in neural information processing systems*, 2007.

[33] O. Contributors, “Openspiel: A framework for reinforcement learning in games,” 2020. arXiv: 1908.09453 [cs.LG].

[34] D. J. Wu, “Accelerating self-play learning in go,” 2020. arXiv: 1902.10565 [cs.LG].

[35] Y. Tian, J. Ma, Q. Gong, S. Sengupta, Z. Chen, J. Pinkerton, and C. L. Zitnick, “Elf opengo: An analysis and open reimplementation of alphazero,” 2019. arXiv: 1902.04522 [cs.AI].

[36] S. Dalton, I. Frosio, and M. Garland, “Accelerating reinforcement learning through gpu atari emulation,” 2020. arXiv: 1907.08467 [cs.LG].

[37] Andy L Jones, “Megastep,” version 0.1, Jul. 7, 2020. [Online]. Available: <https://andyljones.com/megastep>.

[38] J.-B. Grill, F. Altché, Y. Tang, T. Hubert, M. Valko, I. Antonoglou, and R. Munos, “Monte-carlo tree search as regularized policy optimization,” 2020. arXiv: 2007.12509 [cs.LG].

[39] A. Stooke and P. Abbeel, “Accelerated methods for deep reinforcement learning,” 2019. arXiv: 1803.02811 [cs.LG].

[40] S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team, “An empirical model of large-batch training,” 2018. arXiv: 1812.06162 [cs.LG].

[41] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, “Grandmaster level in starcraft ii using multi-agent reinforcement learning,” *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[42] OpenAI, : C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang,

- F. Wolski, and S. Zhang, "Dota 2 with large scale deep reinforcement learning," 2019. arXiv: 1912.06680 [cs.LG].
- [43] K. Cobbe, J. Hilton, O. Klimov, and J. Schulman, "Phasic policy gradient," 2020. arXiv: 2009.04416 [cs.LG].
- [44] PyTorch Contributors, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [45] M. Lezcano-Casado, "Trivializations for gradient-based optimization on manifolds," in *Advances in Neural Information Processing Systems, NeurIPS*, 2019, pp. 9154–9164.
- [46] NumPy Contributors, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: <https://doi.org/10.1038/s41586-020-2649-2>.
- [47] SciPy Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.
- [48] Pandas Contributors, *Pandas-dev/pandas: Pandas 1.0.3*, version v1.0.3, Mar. 2020. DOI: 10.5281/zenodo.3715232. [Online]. Available: <https://doi.org/10.5281/zenodo.3715232>.
- [49] F. Pérez and B. E. Granger, "IPython: A system for interactive scientific computing," *Computing in Science and Engineering*, vol. 9, no. 3, pp. 21–29, May 2007, ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.53. [Online]. Available: <https://ipython.org>.
- [50] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.
- [51] Plotnine Contributors, *Has2kl/plotnine: V0.8.0*, version v0.8.0, Mar. 2021. DOI: 10.5281/zenodo.4636791. [Online]. Available: <https://doi.org/10.5281/zenodo.4636791>.

APPENDIX

A. AlphaZero Implementation

While our implementation was heavily influenced by several different open-source AlphaZero implementations [22], [33]–[35], our unusual use-case - training small agents on small boards - lead to some unusual design decisions.

1) *Small networks*: The original AlphaZero and its open-source replications used very large residual convnets. ELF OpenGo [35], for example, uses a 256-filter 20-block convolutional network, weighing in at roughly 20m parameters and 2 GF-s for a forward pass on a single sample. In our preliminary work however, we found that on the small boards we work with, far smaller - and faster - networks could make it to perfect play.

In particular, we found that perfect play on a 9×9 board can be achieved by a fully-connected residual net with two layers of 512 neurons, along with an input and output layer. This net weighs in at 500k parameters and 500 KF-s for a forward pass, a tiny fraction of the cost of the original AlphaZero networks.

2) *Vectorization*: These very-small networks open the way to further speedups. When the neural networks involved in a reinforcement learning problem are large, the time taken to forward- and backward-propagate through the network dominates the run time of the algorithm. As such, it doesn't often make sense to invest effort in speeding up other parts of the implementation. When the neural networks are small however, these other-parts come to the fore.

In contrast to other AlphaZero implementations, where the environment and tree search are implemented on the CPU, our

implementation is wholly-GPU based. Both the rules of Hex and the tree search codes are written in CUDA and carried out on the GPU. This enables us to massively parallelise things, with a typical training setup collecting experience from 32k Hex boards in parallel.

This is a technique that has been implemented now for a range of environments [36], [37], but ours is the first application of the technique to board games and to MCTS.

3) *Regularized tree search*: If AlphaZero's tree search discovers a particularly high-value strategy during exploration, it can take many, many simulations before the high value of that strategy is fully reflected at the root node. This issue was identified in Grill et al. [38], which also shows it can be resolved by solving a simple optimization problem at each node.

We found that adapting their solution let us use dramatically fewer nodes in our search tree. We did however find that in the small-network regime this work is concerned with, the bisection search proposed by Grill et al. can be a significant factor in the runtime. Fortunately this issue was easily resolved by replacing the bisection search with a Newton search.

We also discovered that the ideal coefficient of exploration, c_{puct} , was far lower than advertised elsewhere in the literature. This corresponds to a lower emphasis on the prior policy distribution versus the findings of the tree search. Our ideal value was in the region of $1/16$, compared to the 2 used in prior work. We remain uncertain as to whether this is due to some peculiarity of our setup, or a consequence of our use of regularized tree search.

4) *Large batches*: It has been observed that many reinforcement learning schemes are substantially faster and more stable when large batch sizes are used [39].

In typical AlphaZero implementations however, the batch size is typically ≈ 1000 samples. We are not certain as to why this is the case, but suspect it is a consequence of large size of the value/policy network limiting how much can be held in memory at once. With our particularly-small networks however, this is much relaxed, and so our runs typically use a batch size of 32k samples. This size was arrived at by calculating the gradient noise scale [40], which is roughly 32k on the largest boards.

5) *Small buffer*: A final discovery was that while many other multi-agent training schemes include large replay buffers [11], tournaments [41] and leagues [42] as a way to suppress cyclic patterns during training, we found that none of these were necessary in our implementation. We do not know if this is a consequence of our choice of game, our small board sizes, our small agents, or our large batches, but the outcome is that we could use a very small replay buffer - 2m samples, or 64 steps of our 32k-replica environment. This lead to a dramatic speedup in training, plausibly due to the much lower staleness of the samples ingested by the learner [42], [43].

6) *Validity*: In all we have made *many* novel adjustments to AlphaZero in this work, and if we were claim superiority over a more conventional implementation then we would be obliged to present a wide array of baselines and ablations.

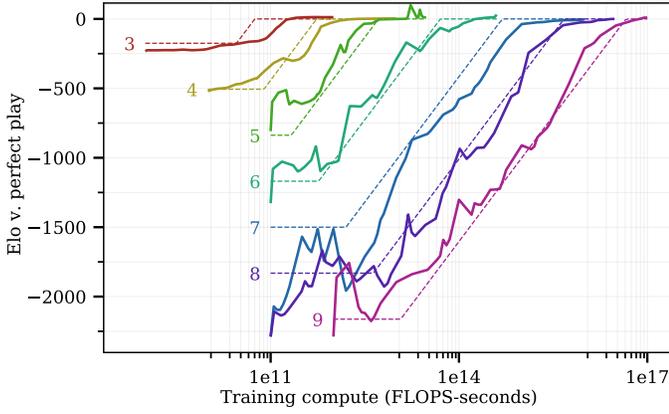


Fig. 10. When computed using top-agent evaluation instead, the frontiers are noisier than in league evaluation but display the same form and similar fits.

TABLE IV
FITTED FRONTIER PARAMETERS (TOP-AGENT EVALUATION)

	m_{flops}	$m_{\text{boardsize}}$	c
plateau		-330	820
incline	600	-590	-4700

However, this work’s goal is not to present a fast Hex solver. The critical property is simply whether our implementation can achieve perfect play, and by comparison to MoHex we find ourselves suitably convinced of this.

B. Handling Non-Transitivity

As discussed in §II-D, Elo ratings are non-transitive. One worry we had was that the compute frontiers observed here might be a product of this non-transitivity and the varying numbers of agents used at different board sizes. To resolve

C. Alternate curve models

We experimented with several functional forms for the compute frontiers.

this worry we also tried evaluating every agent directly against a single top-rated agent.

This ‘top-agent’ evaluation has no transitivity issues, but does require matchups between agents of vastly different skill levels. The 2,000 Elo difference between random and perfect play on a 9×9 board (Fig. 5) implies that the random agent will win 1 in 1m games against the perfect agent. This means we would need to play $\gg 10$ m games to properly resolve the rating of the random agent.

While this is more than we could afford across the 2,800 agents in our stable, we decided to play a limited number of games - 64k - between each agent and a top-rated agent. We found that the frontiers derived using this setup were noisier than the ones generated by playing every agent against every other, but that the pattern was similar in form and fit. The frontiers from this evaluation method can be seen in Fig. 10, and give the fitted parameters in Table IV.

In all, we are convinced that the compute frontiers observed are not due to non-transitivity.

Linear models were our first choice, but the notably non-linear behaviour at the top and bottom of each curve damaged the estimates of the slope of each frontier and the compute required for perfect play.

Sigmoid models meanwhile were much better fits, and their smoothness is arguably a better fit for the phenomena in question. However, that same smoothness makes interpreting their parameters much harder.

The change-point model used in the main text is as good of a fit (in terms of MSE) as the sigmoid model, but its parameters are much easier to interpret.

D. Software

This work depended principally on the libraries pytorch [44], geotorch [45], numpy [46], scipy [47], pandas [48], ipython [49], matplotlib [50] and plotnine [51].