THE UNIVERSITY OF CHICAGO

ARTIFICIAL INTELLIGENCE AND HIGH-PERFORMANCE COMPUTING FOR ACCELERATING STRUCTURE-BASED DRUG DISCOVERY

A DISSERTATION PROPOSAL SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
AUSTIN CLYDE

CHICAGO, ILLINOIS
AUGUST 2022

TABLE OF CONTENTS

ABSTRACT

Traditional techniques for discovering novel drugs are too slow for 21$^{\mathrm{st}}$ challenges, from precision oncology to emerging global pandemics. The COVID-19 Pandemic demonstrated the unequivocal need for rapidly deployable drug discovery capabilities as a matter of national biopreparedness and biosecurity. The challenge is, though, that drug discovery is an immense and complex interdisciplinary field drawing from cheminformatics, bioinformatics, biophysics, machine learning, and high-performance computing among others. To accelerate the screening process of new molecules, researchers are applying developments from artificial intelligence (AI) to the problem; however, the direct application of traditional AI methods overlooks the essential complexities of drug discovery, ranging from protein-conformation flexibility to unique statistical properties of virtual ligand screening. This dissertation presents a unique approach to AI for drug discovery based on tightly integrating insights from biochemistry and biophysics, driving a more accurate and more interpretable drug discovery system, all while leveraging the same accelerating advances from AI. These cross-cutting contributions from AI and HPC workflows illustrate orders of magnitude speedup for computational virtual drug screening, novel workflow designs for high-fidelity screening pipelines which are more accurate than traditional docking, new sampling strategies for exploring novel chemotypes, and complementary workflow analysis techniques which directly links actionable and interpretable goals (such as the design of drugs) with quantitative cost functions. These methodological developments are realized in a case study discovering and validating a novel SARS-CoV-2 3CL-Main Protease inhibitor with a K_i of 2.9 μM (95% CI 2.2, 4.0).

CHAPTER 1

INTRODUCTION

As of August 2022, the COVID-19 pandemic—caused by the SARS-CoV-2 virus—has killed over 6.4 million people worldwide and infected over 8% of the global population (Organization [19]). Within just a few months during the spring and summer of 2022, Monkey Pox—a disease caused by the monkeypox virus from the same genus as variola virus, *Orthopoxvirus*—has already spread globally with over 30,000 cases identified. Even moving beyond the landscape of salient global pandemics, in 2019, it is estimated that between 4 and 6 million people died from disease caused by antimicrobial resistant (AMR) bacteria (Murray et al. [2022]). Beyond pathogens, cancer is a leading killer with over 10 million deaths global. All of these diseases can be solved through small-molecule therapies, in theory.

Yet, the scientific community has only explored a tiny fraction of small molecules. The enumeration and exploration chemicals is no new question: in 1875, Caley published a short note on his enumeration of alkanes utilizing a tree structure (Cayley [1875]). Though Caley's enumeration ended up having a few errors, it is a very early account of treating chemical space as a structured mathematical object (Rains and Sloane [1999]). The design space of small molecule—chemical space—is vast and estimated to consist of about 10^{60} molecules (Bohacek et al. [1996]). Recent research has highlighted that moving beyond the standard chemotypes found in chemsimistry and biology textbooks and lab stocks is extremely fruitful for finding highly potent and selective inhibitors (Jia et al. [2019], Lyu et al. [2022, 2019]). This motivates an immediate need for efficient and automated exploration for synthesis and assay development for various applications, including drug discovery and materials design. Computational enumeration of chemical space is a long-studied problem since the early ages of computing (Cernak [2018]). The current state of the art projects have enumerated around 2 billion drug-like compounds, and GDB has around 166 billion compounds of up to 17 atoms of C, N, O, S, and halogens (Patel et al. [2020], Ruddigkeit et al. [2012]).

Given this unimaginably dense design space, the question is *how to identify drugs which can, with the help of other medical countermeasures and social controls, prevent disease, be personalised to reduce side-effects and increase efficacy, and be readily available cheaply, in real-time, and globally distributed*? No pharmaceutical company may share all of these goals, but there are academic endeavours which have shown promise both in the scientific and social axis of this vision. During the COVID-19 pandemic, two academic group have identified small-molecular inhibitors such as the National Virtual Biotechnology Laboratory and the COVID Moonshot Project (Clyde et al. [2021b], Achdout et al. [2020]). Both groups were able to identify lead compounds within 9 months of the beginning of the pandemic. Although the transition from drug lead to downstream toxicity, animal model, and human studies is a challenging and more difficult road (a open area for improvement—no doubt), this feat should invite further perspective and analysis on how this was possible?

These two groups, of which I was entrenched in both, pursued approaches to drug discovery unlike those found in tightly sealed corporate research and development offices. They focused on leveraging a large and interdisciplinary community of scientists from private, public, and academic laboratories (Clyde [2022a]). They pursued computational techniques capable of leveraging the United States Department of Energy supercomputing infrastructure (Buchanan and Streiffer [2020], Bhati et al. [2021]) and, globally, the extra cycles of personal computers forming an decentralized "exascale" supercomputer (Zimmerman et al. [2021]). Lastly, they were both founded on the ideal of fully open and global science. Computational drug discovery has great potential for democratizing an industry facing many societal pressures.

In this dissertation, I will focus on the technical developments to accelerate and expand these drug discovery programs, particularly in AI and high-performance computing (HPC). In the following chapters, I will outline new approaches to scaling and applying these methods as part of a broader goal. I envision these systems as part of a basic biosecurity surveillance program capable of monitoring the environment, identifying threats, and queuing wet-lab automated testing so that when epidemiology officials declare a disease spreading uncontrollable, a treatment already is known and ready.

Artificial intelligence (AI) is a form of computational problem solving that a aims to (1) mimic some human notion of intelligence (whether it be the ability to naturally use language or identify objects in the world) and (2) scale these capabilities to process vast data outside the window reasonable by a single person. Recently, advances in deep learning have revitalized AI's research program through breakthroughs in natural language processing (Brown et al. [2020]), game-playing (Silver et al. [2016]), protein-folding (Cramer [2021]), code-generation (Chen et al. [2021]), and image-classification (He et al. [2016], Chai et al. [2021]). Although none of these models represent paradigmatic solutions to the problem[1], these models are, at a minimum, surprisingly adept and capable at convincing one that they are nearing solutions. In particular engineering domains, scientific modeling is increasingly merging with the technical and structural ideas presented in AI papers broadly encapsulating the idea of "AI for science" (Stevens et al. [2020]). If modeling offers little scientific insight from a theoretical perspective, what then can we make of AI for science (AI4S)?

One can raise the question that the science in the ideas percolating around AI4S might be a different kind of science. This dissertation is by no means an attempt to characterize the present status of computational scientific research nor fortune-tell its future; however, it is certainly a case study in AI4S. I will offer at moments a perspective on the broader ambition on how one can connect semantic meaning such as that sought in traditional science with the often overly-used quip that deep learning is a black box.[2] There are theoretical and

[1] A paradigmatic solution would most likely come with some theory fitting the present ideas around what it means to solve a scientific challenge. None of these models present any theoretical breakthroughs within the terms of the problems they attempt to solve. Furthermore, many of these models require great set up, specification, and are limited in scope within those problem areas.

[2] See forthcoming article "Why You Were Told Deep Learning is a Black Box: On Expertise and Politics in Scientific Practice".

practical assumptions carried with computational tools, and, similarly, the computational frame—a kind of bias required to force a scientific problem through the structure of a AI problem—does impact, affect, and alter the science (Anderson [2008]). This co-production of the sciences and its theory alongside AI4S, its algorithmic assumptions, and the tools will be directly discussed at various moments (in particular in Chapter 9) (Jasanoff [2004]).

Given the interdisciplinary required for approaching computational drug discovery, the first two chapters of this dissertation outline background information for computer science minded researchers to understand the intricacies for drug discovery. Chapter 2 provides an interdisciplinary overview of the biology and biophysical background for drug discovery. While the chapter mainly aimed to review concepts such as different models of induced-fit protein-ligand binding and free-energy perturbation methods, it aims to concisely connect a computationally-minded reader into the various theoretical levers biophysics and biology have to offer.

Chapter 3 presents an overview of virtual ligand screening (VLS) as a computational challenge. Recent advances in computational screening techniques are outlined, including an overview of the different paradigms of drug discovery such as ligand and structure based drug discovery. Chapter 3 presents a contribution to ligand based drug design for precision oncology while the rest of this dissertation is devoted to structure-based drug design.

Chapter 4 outlines the different workflow designs for AI in HPC drug discovery workflows with a focus applying AI to structure based drug design. It begins with the simple idea of using surrogate models to replace slow and less-accurate code from traditional VLS workflows. This small building block is connected with different design ideas ranging from generative modeling to database searches.

Chapter 5 demonstrates precisely how chemical structures are can be used in deep neural network (DNN) based surrogate models. The challenge for training surrogate models is that the standard statistical assumptions stable neural network training relies on are not applicable for the highly-skewed distributions in drug discovery. This chapter showcases different techniques that can be applied to neural network training for improving the accuracy of speed up and illustrates various trade-offs of different chemical featurization techniques.

Chapter 6 showcases how increasingly accurate and more costly computational simulation techniques can be combined with surrogate models to achieve more accurate screening workflows without increasing the overall computational cost.

Chapter 7 removes the assumption that virtual screening problems maintain no structure over the screening space by arguing that chemical space can be thought of as a multi-hypergraph partitioned by basic theoretical tools from cheminformatics such as Bemis-Murcko scaffold decomposition and electrostatics.

Chapter 8 presents new workflow evaluation metrics which couple business decisions such as spending for purchasing compounds directly with model performance and high-performance computing measures. Together, this presents a complete model of exactly how a model's performance is coupled to throughput and hit-identification as well as how computational characteristics of an HPC platform will scale those metrics.

Finally, chapter 9 summarizes various opportunities for future work such as utilizing advances in large-language models for improving the transferability of disparate data sets, improving the calibration and uncertain quantification, connecting active learning and automatic laboratories with drug discovery workflows, and finally generative drug design techniques.

CHAPTER 2

DRUG DISCOVERY: INTRODUCTION

Summary:

This chapter will outline drug discovery with a focus on the biological components. The first section will overview the basics of disease caused by pathogens and cancer demonstrating the need for small molecules. The second section will address the biochemistry such as proteins, their flexibility, binding pockets, and the induced-fit model of drug binding. The third section will characterize drug discovery efforts such as the basics of screening, dose response curves, and quantitative structure activity studies (Kneller et al. [2021]). The last section will highlight computational efforts besides virtual screening such as ligand-based search, large scale simulation (Dommer et al. [2021], Casalino et al. [2021]), and cheminformatics featurization (Babuji et al. [2020]).

CHAPTER 3

VIRTUAL SCREENING: BACKGROUND AND HISTORY

Summary:

This chapter will focus directly on the topic of the remaining dissertation: virtual screening. The first section will cover some basics of computational screening broadly (not just for drugs). The second section will cover the history of virtual drug screening and the predominant ideologies (structure-based drug design and ligand-based drug design). This section will also feature some of my work in ligand-based drug design in particular for personalized cancer therapy. This will only be a minor example as this dissertation is focused on structure based drug design (Xia et al. [2022], Partin et al. [2021], Clyde et al. [2020c,a]). The third section will focus on protein-ligand docking and the computational details of it (such as binding site identification, conformer generation, scoring function, available tools, and their computational performance characteristics).

CHAPTER 4

AI AND VIRTUAL SCREENING HPC WORKFLOWS

Summary:

This chapter will focus on how AI can be introduced into the basic VLS pipeline introduced in the previous chapter through the concept of surrogate models. This will include details on modeling as a data science problem protein-ligand docking, the kinds of workflows that are possible with this application of AI, and basic presentation of the various models used (drug descriptors, graph models, RNNs, smiles, images, etc.). This chapter will be primarily an extension of my book chapter on AI for high-throughput drug screening (Clyde [2022b]) and my chapter on large language models for science AI4S workflows (*forthcoming*).

CHAPTER 5

SURROGATE MODELS FOR ACCELERATED DOCKING

Summary:

This chapter will focus on my particular modeling efforts in building surrogate models for docking. This chapter will focus on the performance characteristics of docking with a focus on the COVID-19 inhibitor discovery (Clyde et al. [2021a], Wu et al. [2021]). I will outline the datasets required, the models developed, the scale of their deployment, scaling characteristics, and error characteristics.

CHAPTER 6

TIERED-WORKFLOWS

Summary:

This chapter will address my work with tiered-workflows—where cheap but inaccurate surrogates are used to screen compounds, sending a smaller subset to a more expensive but also more accurate kernel (such as simulation). The pipeline I will discuss is called IMPECCABLE Saadi et al. [2021], Bhati et al. [2021], Clyde et al. [2019]. The first section will introduce the idea, the second section will outline the theoretical considerations (extending the mathematics from Woo et al. [2021], the third section will outline the results of the pipeline with some of the COVID-19 molecules, and finally the last section will discuss the performance charateristics.

CHAPTER 7

SAMPLING STRATEGIES FOR CHEMICAL SPACE

This chapter will address the bootstrapping and sampling problem: how do you select what subset of molecules to get an initial dataset for modeling building on? Even if you have a model built, how do you sample chemical space given you cannot access the whole thing? This chapter will present the idea of using a hypergraph along different chemical-theoretic axes such as scaffolds and pharmacophores. This chapter will comment on the use of LLMs for storing this graph and accessing it on the fly (Clyde et al. [2021c]).

CHAPTER 8

ANALYSIS OF VIRTUAL SCREENING MODELS

Summary:

This chapter will overview how to analyze virtual screening models. The first section will be historical regarding the different sets of metrics used in the virtual screening community and their incongruence with those used in the deep learning community. The second section will introduce the idea of regression enrichment surfaces (Clyde et al. [2020b]). The third section will relate to this computational scaling (Lee et al. [2021], Ma et al. [2018]).

CHAPTER 9

CONCLUSION AND OPPORTUNITIES

Summary

In this chapter I will conclude with summarizing the developments in the previous chapters. I will also comment more broadly on different strategies that one can use beyond the HPC/AI workflows outlined in this chapter such as generative modeling and coupling it with simulations (RLMM). The third section will offer some forward looking ideas about how the next order of magnitude speeds up can be sought. The final section will conclude with ideas around automated discovery and autonomous laboratories.

REFERENCES

Hagit Achdout, Anthony Aimon, Elad Bar-David, Haim Barr, Amir Ben-Shmuel, James Bennett, Melissa L Bobby, Juliane Brun, BVNBS Sarma, Mark Calmiano, et al. Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *BioRxiv*, 2020.

Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

Yadu Babuji, Ben Blaiszik, Tom Brettin, Kyle Chard, Ryan Chard, Austin Clyde, Ian Foster, Zhi Hong, Shantenu Jha, Zhuozhao Li, et al. Targeting sars-cov-2 with ai-and hpc-enabled lead generation: a first data release. *arXiv preprint arXiv:2006.02431*, 2020.

Agastya P Bhati, Shunzhou Wan, Dario Alfè, Austin R Clyde, Mathis Bode, Li Tan, Mikhail Titov, Andre Merzky, Matteo Turilli, Shantenu Jha, et al. Pandemic drugs at pandemic speed: infrastructure for accelerating covid-19 drug discovery with hybrid machine learning-and physics-based simulations on high-performance computers. *Interface focus*, 11(6):20210018, 2021.

Regine S Bohacek, Colin McMartin, and Wayne C Guida. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, 16(1): 3–50, 1996.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Michelle V Buchanan and Stephen Streiffer. Nvbl (national virtual biotechnology laboratory) overview. Technical report, USDOE Office of Science (SC)(United States), 2020.

Lorenzo Casalino, Abigail C Dommer, Zied Gaieb, Emilia P Barros, Terra Sztain, Surl-Hee Ahn, Anda Trifan, Alexander Brace, Anthony T Bogetti, Austin Clyde, et al. Ai-driven multiscale simulations illuminate mechanisms of sars-cov-2 spike dynamics. *The International Journal of High Performance Computing Applications*, 35(5):432–451, 2021.

E Cayley. Ueber die analytischen figuren, welche in der mathematik bäume genannt werden und ihre anwendung auf die theorie chemischer verbindungen. *Berichte der deutschen chemischen Gesellschaft*, 8(2):1056–1059, 1875.

Tim Cernak. A machine with chemical intuition. *Chem*, 4(3):401–403, 2018.

Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Austin Clyde. Ai for science and global citizens. *Patterns*, 3(2):100446, 2022a.

Austin Clyde. Ultrahigh throughput protein–ligand docking with deep learning. In *Artificial Intelligence in Drug Design*, pages 301–319. Springer, 2022b.

Austin Clyde, Dave Wright, and Shantenu Jha. Cafcw 120 integrating high-performance simulations and learning toward improved cancer therapy. 2019.

Austin Clyde, Tom Brettin, Alexander Partin, Maulik Shaulik, Hyunseung Yoo, Yvonne Evrard, Yitan Zhu, Fangfang Xia, and Rick Stevens. A systematic approach to featurization for cancer drug sensitivity predictions with deep learning. *arXiv preprint arXiv:2005.00095*, 2020a.

Austin Clyde, Xiaotian Duan, and Rick Stevens. Regression enrichment surfaces: a simple analysis technique for virtual drug screening models. *arXiv preprint arXiv:2006.01171*, 2020b.

Austin Clyde, Arvind Ramanathan, and Rick Stevens. Virtual screening with deep learning using cancer cell line dose-response data. *Clinical Cancer Research*, 26 (12_Supplement_1):36–36, 2020c.

Austin Clyde, Thomas Brettin, Alexander Partin, Hyunseung Yoo, Yadu Babuji, Ben Blaiszik, Andre Merzky, Matteo Turilli, Shantenu Jha, Arvind Ramanathan, et al. Protein-ligand docking surrogate models: A sars-cov-2 benchmark for deep learning accelerated virtual screening. *arXiv preprint arXiv:2106.07036*, 2021a.

Austin Clyde, Stephanie Galanie, Daniel W Kneller, Heng Ma, Yadu Babuji, Ben Blaiszik, Alexander Brace, Thomas Brettin, Kyle Chard, Ryan Chard, et al. High-throughput virtual screening and validation of a sars-cov-2 main protease noncovalent inhibitor. *Journal of chemical information and modeling*, 62(1):116–128, 2021b.

Austin Clyde, Bharat Kale, Maoyuan Sun, Michael Papka, Arvind Ramanathan, and Rick Stevens. Scaffold embeddings: Learning the structure spanned by chemical fragments, scaffolds and compounds. In *Workshop on Learning Meaningful Representation of Life*, 2021c.

Patrick Cramer. Alphafold2 and the future of structural biology. *Nature Structural & Molecular Biology*, 28(9):704–705, 2021.

Abigail Dommer, Lorenzo Casalino, Fiona Kearns, Mia Rosenfeld, Nicholas Wauer, Surl-Hee Ahn, John Russo, Sofia Oliveira, Clare Morris, Anthony Bogetti, et al. # covidisairborne: Ai-enabled multiscale computational microscopy of delta sars-cov-2 in a respiratory aerosol. *bioRxiv*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Sheila Jasanoff. The idiom of co-production. In *States of knowledge*, pages 1–12. Routledge, 2004.

Xiwen Jia, Allyson Lynch, Yuheng Huang, Matthew Danielson, Immaculate Lang'at, Alexander Milder, Aaron E Ruby, Hao Wang, Sorelle A Friedler, Alexander J Norquist, et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573(7773):251–255, 2019.

Daniel W Kneller, Hui Li, Stephanie Galanie, Gwyndalyn Phillips, Audrey Labbé, Kevin L Weiss, Qiu Zhang, Mark A Arnould, Austin Clyde, Heng Ma, et al. Structural, electronic, and electrostatic determinants for inhibitor binding to subsites s1 and s2 in sars-cov-2 main protease. *Journal of medicinal chemistry*, 64(23):17366–17383, 2021.

Hyungro Lee, Andre Merzky, Li Tan, Mikhail Titov, Matteo Turilli, Dario Alfe, Agastya Bhati, Alex Brace, Austin Clyde, Peter Coveney, et al. Scalable hpc & ai infrastructure for covid-19 therapeutics. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pages 1–13, 2021.

Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

Jiankun Lyu, John Irwin, and Brian Shoichet. Modeling the expansion of virtual screening libraries. 2022.

Heng Ma, Austin Clyde, Anda Trifan, Venkatram Vishwanath, Arvind Ramanathan, Debsindhu Bhowmik, and Shantenu Jha. Benchmarking machine learning workloads in structural bioinformatics applications. *interactions*, 27:32, 2018.

Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022.

World Health Organization. Who covid-19 dashboard, 19.

Alexander Partin, Thomas Brettin, Yvonne A Evrard, Yitan Zhu, Hyunseung Yoo, Fangfang Xia, Songhao Jiang, Austin Clyde, Maulik Shukla, Michael Fonstein, et al. Learning curves for drug response prediction in cancer cell lines. *BMC bioinformatics*, 22(1):1–18, 2021.

Hitesh Patel, Wolf-Dietrich Ihlenfeldt, Philip N Judson, Yurii S Moroz, Yuri Pevzner, Megan L Peach, Victorien Delannée, Nadya I Tarasova, and Marc C Nicklaus. Savi, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Scientific data*, 7(1):1–14, 2020.

Eric M Rains and Neil JA Sloane. On cayley's enumeration of alkanes (or 4-valent trees). *Journal of Integer Sequences*, 2:Art–No, 1999.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Aymen Al Saadi, Dario Alfe, Yadu Babuji, Agastya Bhati, Ben Blaiszik, Alexander Brace, Thomas Brettin, Kyle Chard, Ryan Chard, Austin Clyde, et al. Impeccable: integrated modeling pipeline for covid cure by assessing better leads. In *50th International Conference on Parallel Processing*, pages 1–12, 2021.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Rick Stevens, Valerie Taylor, Jeff Nichols, Arthur Barney Maccabe, Katherine Yelick, and David Brown. Ai for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States), 2020.

Hyun-Myung Woo, Xiaoning Qian, Li Tan, Shantenu Jha, Francis J Alexander, Edward R Dougherty, and Byung-Jun Yoon. Optimal decision making in high-throughput virtual screening pipelines. *arXiv preprint arXiv:2109.11683*, 2021.

Yulun Wu, Nicholas Choma, Andrew Chen, Mikaela Cashman, Érica T Prates, Manesh Shah, Verónica G Melesse Vergara, Austin Clyde, Thomas S Brettin, Wibe A de Jong, et al. Spatial graph attention and curiosity-driven policy for antiviral drug discovery. *arXiv preprint arXiv:2106.02190*, 2021.

Fangfang Xia, Jonathan Allen, Prasanna Balaprakash, Thomas Brettin, Cristina Garcia-Cardona, Austin Clyde, Judith Cohn, James Doroshow, Xiaotian Duan, Veronika Dubinkina, et al. A cross-study analysis of drug response prediction in cancer cell lines. *Briefings in bioinformatics*, 23(1):bbab356, 2022.

Maxwell I Zimmerman, Justin R Porter, Michael D Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L Mallimadugula, Catherine E Kuhn, Jonathan H Borowsky, Rafal P Wiewiora, et al. Sars-cov-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature chemistry*, 13(7):651–659, 2021.