Adversarial Attacks against AI-driven Experimental Peptide Design Workflows*

Arvind Ramanathan
Data Science and Learning Division
Argonne National Laboratory
Lemont IL 64309, USA
ramanathana@anl.gov

Sumit Kumar Jha
Computer Science Department
University of Texas at San Antonio
San Antonio TX 78249, USA
sumit.jha@utsa.edu

***Abstract*—Artificial intelligence and machine learning (AI/ML) techniques are fueling a revolution in how scientific experiments are designed, implemented and automated. Specifically, increasing high-bandwidth instruments coupled to new hardware and software systems can significantly improve the throughput of experimental results, while AI/ML techniques can provide insights into novel science and theories that were hitherto inaccessible. Despite recent progress in such "self-driving labs", these automated platforms are susceptible to adversarial attacks as well as more traditional cybersecurity attacks. Using a motivating example of an automated approach to design anti-microbial peptides (AMP), our position paper seeks to demonstrate how a lack of adversarial robustness of AI systems such as protein folding networks may affect the execution of such experimental workflows. We highlight important problems in adversarial robustness that may need to be resolved in order to establish a trustworthy and safe AI-driven AMP synthesis system.**

***Index Terms*—AMP synthesis, adversarial attacks, protein folding networks, safety and trustworthy**

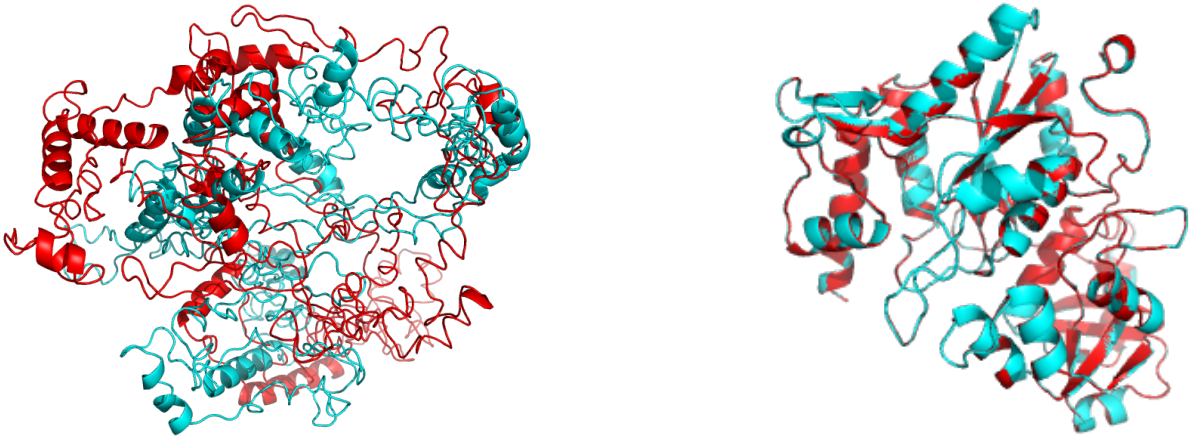
I. Introduction

Artificial intelligence (AI) and machine learning (ML) techniques are promising to revolutionize how experiments are designed, implemented and executed [1]. Already several examples of automated platforms for scientific experiments exist, where robotic instrumentation and AI/ML techniques have successfully executed end-to-end scientific workflows [2]–[6]. While such examples have proliferated materials science [7]–[9], synthetic biology [10], [11] and chemistry/drug-discovery applications [12], widespread deployment of such AI-enabled automated platforms remains quite challenging [13]. These challenges can be largely attributed to: (1) diversity in the deployment of robotic instrumentation platforms across the scientific enterprise; (2) lack of well defined standards for development and exchange of experimental protocols and data; (3) intrinsic difficulty in connecting diverse instruments with heterogeneous computing infrastructures and (4) emerging adversarial robustness [14]–[16] and cybersecurity concerns in coupling physical instruments with AI/ML approaches.

Funding for this work was provided by the Laboratory Directed Research and Development project at Argonne National Laboratory. The research was partially supported by the National Science Foundation SPX award #1822976, National Science Foundation award #2113307, ONR grant #N00014-21-1-2332, and the DARPA co-operative agreement #HR00112020002. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Fig. 1. Robotic manipulation for automating the synthesis of antimicrobial peptides (AMPs) using AI/ML approaches.

The primary focus of our position paper is in outlining some of the cybersecurity and adversarial robustness concerns when deploying automated experimental platforms. We focus on an illustrative workflow involving the design of small peptides exhibiting antimicrobial activity against environmental *Escherichia coli* (E. coli) samples. However, these observations can be generalized to other large-scale experimental design workflows where artificial intelligence or machine learning techniques interact with and potentially 'execute' experimental commands through such robotic instruments. We briefly recall recent results in adversarial robustness of protein folding networks and illustrate adversarial attacks on the RoseTTAFold network [17] using naturally occurring proteins as well as antimicrobial peptides.



(i) RMSD=34.162Å (ii) RMSD=0.119Å

Fig. 2. Adversarial attacks against the RoseTTAFold network obtained from our earlier work [18]. The predicted structure for the original (shown in blue) and the adversarial (shown in red) protein sequences for the SFP1 interferon stimulator (left) and 2QJF_1 human bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthetase 1 (right) as predicted by RoseTTAFold and aligned using PyMOL. RMSD is the Root Mean Square Distance, typically measured in Å (10^{-10}m). The predicted structure on the right is robust to biologically small perturbations, while the one on the left is too sensitive even to biologically small changes.

II. AI-driven design of protein/peptides with antimicrobial activity

Anti-Microbial Peptides (AMP) provide a new frontier in defending our health and well-being against pathogens such as bacteria and viruses. AMPs can influence cellular membranes [19] and then affect cellular processes. Recently, AI-methods have been used in designing novel AMPs targeting various bacterial species [20].

An automated pipeline to facilitate the screening of AMPs for pathogen activity is based on a 4 step process. *First*, we employ protein language models and generative models to explore the space of possible AMP designs. Popular language models like BERT [21] have been fine-tuned on protein sequences and the resulting ProBERT model can be used to generate candidate AMP sequences. Such models are usually easier to train but can still be used to predict antimicrobial behavior using experimental feedback. Generative models such as variational encoders [22] and Wasserstein encoders [20] can also be used to embed the space of AMPs into a latent space, and then sample from this latent space to create novel AMPs. Features such as length and charge distribution of AMPs can be sampled from the latent space and correlated with experimental observations [20].

Second, we investigate the impact of candidate AMPs on membranes using molecular dynamics simulations and employ AI/ML methods to predict protein-membrane interactions. A machine learning library for analyzing molecular dynamics simulations, such as mdlearn [23] is then used to predict MD outcomes. Figure 3 shows snapshots of an AMP protein interaction with a membrane.

Third, we automate the physical testing of the designed AMP using an experimental assay that quantifies the effective killing of the bacterial population based on the concentration of the peptide. This requires the ability to program robotic manipulators that can move liquids around and read plates for absorbance. The automated observation of the experiments also requires the design of customized vision models in visible and possibly other spectra.

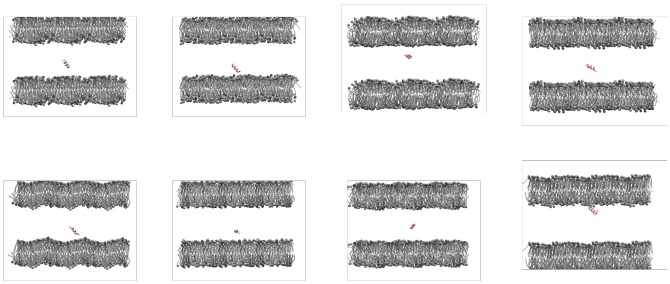


Fig. 3. Molecular dynamics simulations of an anti-microbial protein (AMP) interacting with the cell membrane. End-to-end differentiable AI/ML models for predicting such dynamics may be susceptible to adversarial attacks.

Fourth, we will use artificial intelligence methods to infer activity of designed AMPs against different microbial strains and predict which residues should be modified for the next round of experimental investigations. Each of the steps are susceptible to cybersecurity attacks as well as a lack of adversarial robustness, which we describe below.

III. Adversarial Robustness

Besides traditional cybersecurity attacks, AI-driven design of proteins/peptides can also be subjected to adversarial attacks either due to lack of robustness of models or due to models with Trojans injected into them. We briefly survey a few different kinds of models employed in AI-driven design of AMPs and the need for adversarial robustness in these models.

1) Attacks on BERT and ProteinBERT models: Adversarial attacks on two fundamental natural language tasks, text classification and textual entailment, has been successfully

illustrated on the powerful pre-trained BERT model using the TEXTFOOLER [16] system. It has been shown that the BERT model is not even adversarially robust to misspellings [24].

ProteinBERT is a fine-tuned variant of BERT that can predict protein structures, biophysical attributes and post-translational modifications from protein sequences using relatively small amounts of training data. ProteinBERT may be susceptible to the same sort of adversarial attacks that fool traditional BERT systems. This poses a challenge for protein synthesis because predictions from ProteinBERT may be susceptible to small perturbations in the input.

2) Attacks on Protein AutoEncoders: Adversarial attacks on autoencoders have been illustrated on images and on machine communications. Attacks on variational autoencoders and conditional variational autoencoders have been demonstrated on multiple image data sets, including MNIST, SVHN and CelebA [25]. Adversarial attacks on physical wireless signals for machine to machine communication has been shown to be more damaging than merely jamming the wireless transmission [26].

It is not known if latent space representations of proteins are robust to adversarial perturbations. Protein autoencoders are likely to be susceptible to adversarial attacks, and small perturbations in the input may lead to different predictions.

3) Attacks on Protein Folding Neural Networks: Protein folding neural networks like AlphaFold and RoseTTAFold perform very well in performing three-dimensional structures of protein from their amino acid sequences. However, it has been recently shown that RoseTTAFold is susceptible to adversarial attacks [18]. It is likely that the current generation of protein folding neural networks are all susceptible to adversarial examples to varying degrees. The creation of adversarially robust protein folding neural networks remains a challenge.

Recent work [18] has demonstrated the susceptibility of RoseTTAFold to adversarial attacks [27] by generating several examples where protein sequences that vary only in five residues result in very different three-dimensional protein structures. The approach employs sequence alignment scores [28] such as those derived from Block Substitution Matrices (BLOSUM62) to identify a space of biologically similar protein sequences used in constructing adversarial perturbations. Computational experimental studies show that different input protein sequences have very different adversarial robustness. In this study, the RMSD in the protein structure predicted by RoseTTAFold [17] ranges from 0.119Å to 34.162Å when the adversarial perturbations are bounded by 20 units in the BLOSUM62 distance.

4) Attacks on Vision Models: Manipulation of fluids by experimental robots need to be observed by computer vision algorithms and may be used to drive end-to-end deep learning algorithms for control. Figure 4 shows an example of a rather intricate vision classification challenge posed by the peptide synthesis process. The robotic manipulation of the synthesized peptide depends on the type of the synthesized protein – soluble, pellets or hairgel.

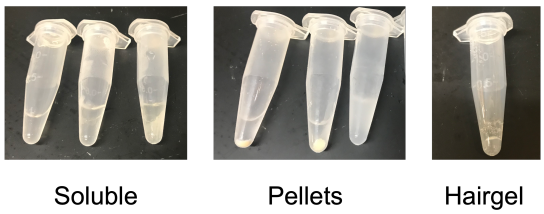


Fig. 4. Different forms of proteins such as soluble, pellets and hairgel obtained during synthesis of peptides.

Fig. 5. Illustration of a patch attack on an image.

Adversarial examples in computer vision have been known for both neural networks [14], [15] and traditional machine learning methods such as support vector machines [29]. Figure 5 shows an example of a patch attack on an image; a small patch causes the image to be labeled as a hook instead of a brambling bird. Hence, it is likely that vision algorithms used in end-to-end control of protein synthesis robots may suffer from similar adversarial patch attacks that may not be prevented by traditional cybersecurity protections.

Several approaches for building robustness in neural networks have been investigated. For example, it has recently been shown that attributions of neural SDE models are more robust [30] than those obtained from traditional residual network models. See Fig. 6 for an illustration of explanations obtained using this method. Other approaches including adversarial training have also been used to train robust neural networks.

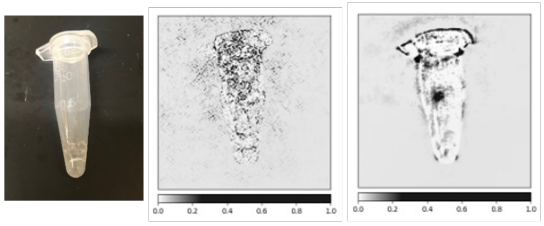
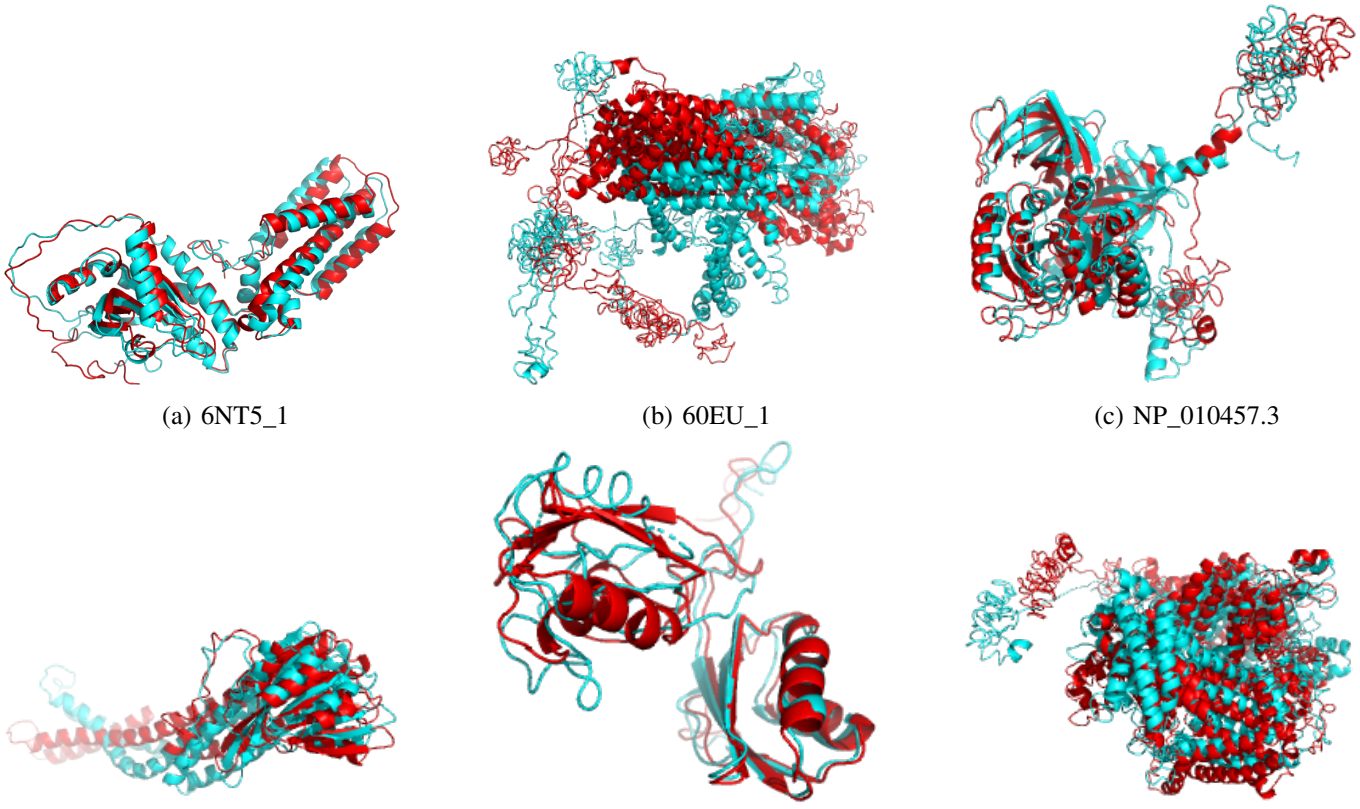


Fig. 6. Visualization of the attribution obtained by a non-robust model (center) and a robust model (right) on the input image (left).

Several challenges remain in the application of adversarial robustness to proteins. First, the distance measures between the input sequences and the metrics for specifying robustness of predicted outputs have not been well-studied for adversarial attacks on proteins. Second, robustness of ProteinBERT and related models need to be biologically inspired so that biologically significant changes lead to substantial changes in predictions. Third, image recognition tasks in the wet lab are more subtle that the challenging ImageNet benchmark; hence, robustness and attribution analysis for such lab-based vision benchmarks need to be investigated.



(a) 6NT5_1 (b) 60EU_1 (c) NP_010457.3

(d) 2BBJ_1 (e) 2ROP_1 (f) 6NT3_1

Fig. 7. Adversarial attacks against the RoseTTAFold network obtained from our earlier work [18]. The predicted structure of protein sequences (blue) and the structures of their adversarial perturbations (red) produced using RoseTTAFold and aligned using PyMOL.

IV. Experimental Results

We briefly survey results from our recent work on adversarial robustness for protein folding networks [18] and present new results on the robustness of anti-microbial peptides.

A. Adversarial Robustness for Naturally Occurring Proteins

We recall the results of adversarial attacks on 6 naturally occurring proteins shown in Fig. 7. The adversarial sequences and the original sequences differed in BLOSUM62 distance of at most 20, and are thus biologically close to each other.

1) 6NT5_1: Stimulator of interferon protein in human beings: As shown in Fig. 7(a), the alignment involving all residues has a RMSD score of 5.026Å showing that the adversarial structure is quite different.

2) 6OEU_1: Protein patched homolog 1 : Figure 7(b) shows the aligned structures for the sequence and its adversarial perturbation. The two sequences were aligned using PyMOL and achieved a RMSD of 34.162Å.

3) NP_010457.3: Yeast translation termination factor GTPase eRF3: Figure 7(c) shows the structure of the original sequence and the structure of the adversarial sequence aligned together with a high RMSD of 6.870Å.

4) 2BBJ_1: Eubacteria CorA Mg2+ transporter: Figure 7(d) shows the structures for the original and adversarial sequences with a high RMSD score of 6.76Å.

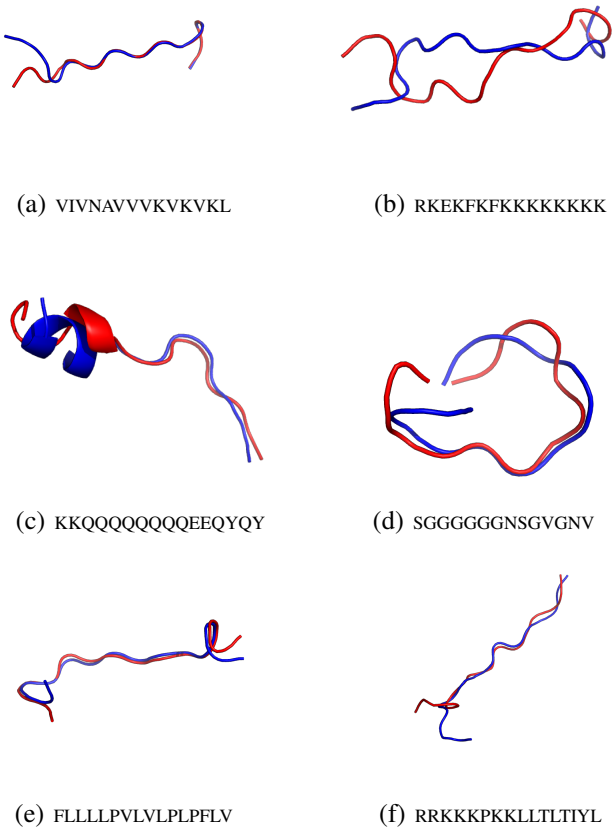
5) 2ROP_1: Human Copper-transporting ATPase 2: As illustrated in Figure 7(e), the alignment between the structure corresponding to this sequence and its adversarial perturbation has a high RMSD score of 8.495.

B. Adversarial Robustness for Anti-Microbial Peptides

We investigate the robustness of RoseTTAFold protein structure predictions to adversarial attacks on 5 different AMP sequences to understand how small changes in the AMP sequence can cause changes in the structure of the molecules.

Figure 8 shows the structure of the original sequence in blue and the adversarial sequence in red aligned together using PyMOL. The adversarial sequence is obtained by changing only 2 amino acids in the sequence of the protein and attacking the end-to-end neural network of RoseTTAFold.

The existence of such adversarial sequences whose structures are different from the input sequence despite the sequence similarity leads to two interesting possibilities: (i) The neural network may not be robust and may produce different outcomes even on proteins where we do not anticipate structural changes. (ii) The protein structures for these AMPs may be inherently fragile and small changes may lead to great structural diversity in AMPs. Further experimental evaluation of the adversarial structures may resolve which of these possibilities is true in practice.



(a) VIVNAVVVKVKVKL (b) RKEKFKFKKKKKKKK

(c) KKQQQQQQQQEEQYQY (d) SGGGGGGNSGVGNV

(e) FLLLLPVLVLPLPFLV (f) RRKKKPKKLLTLTIYL

Fig. 8. Adversarial examples for AMPs using RoseTTAFold.

V. Conclusions

Our position paper highlights the emerging cybersecurity and adversarial robustness concerns of using AI-driven experimental loops. Unlike traditional cybersecurity attacks, AI/ML methods can be challenged by seemingly innocuous objects like well-designed patches printed on lab equipment. Such patch attacks may not be prevented by traditional security or cybersecurity methodologies and may require new innovative protection protocols. Possibilities including the design of deep neural network models that can scan the environment and detect suspicious patches in a laboratory setup.

The injection of Trojans during the training of protein folding networks can lead to scenarios where a protein folding network that perform well on naturally occurring proteins but deliberately make incorrect predictions on a subset of antimicrobial proteins (AMPs). Such incorrect predictions can prevent the discovery of clinically significant peptides without the end user realizing the deliberate sabotage introduced into the model. Such Trojans could be injected into the AI/ML models by suitably modifying public databases usually employed by scientists. Trojan models may also be created by launching a cybersecurity attack that does not cause any other easily perceptible effects.

Our analysis of the robustness of protein folding neural networks has only focused on change in predictions under biologically small perturbations. In future efforts, we seek to validate these against ground truth predictions.

The ubiquitous and often seamless continuity between scientific instruments and high performance computing infrastructure means that there is a tremendous opportunity to unveil new scientific theories and benefit mankind in multitude of ways. However such gains need to be balanced with a critical view of the intrinsic cyber-vulnerability of these systems. With internet of things (IoT) and the development of advanced sensor networks, the intrinsic vulnerabilities of such connected systems poses immense cybersecurity challenges that are yet to be fully understood [31].

An attack on a closed-loop AI-driven workflow is likely to be more obstructive and harder to detect than the influence of the Stuxnet worm on the operation of Iranian nuclear reactors [32]. As shown in our adversarial attack examples in Sec. IV, a cyberattack that carefully changes a couple of amino acids during synthesis of peptides would produce structurally different proteins. Subsequent experimental evaluation of these perturbed peptides and their inclusion in the design of experiments would lead to severe system degradation and probably a complete failure of the AI-driven peptide design without any significant visibility of the attack to human scientists. Thus, novel approaches for detecting such cyberattacks on AI-driven design systems need to be investigated.

Robustness of neural networks have been widely investigated over the last decade [14], [33]–[35]. However, AI-driven experiment design creates novel challenges for robustness in artificial intelligence. As shown in Fig. 4, neural networks that monitor experimental setups for rare unforeseen outcomes and actively seek human intervention may need to be developed. Robust detection of out-of-distribution (OOD) data [36] and experimental setups would be crucial in ensuring continuous and smooth functioning of AI-driven workflows.

The communication between AI and human experts is also crucial to enhance the synergy between man and machine. In particular, interpretation methods that explain neural network decisions to end users would be crucial in understanding, diagnosing and repairing end-to-end AI driven design workflows in cases of failure. While interpretation methods for image classification tasks have been substantially developed [30], AI explanation methods for time-series data and language-like models are less understood.

VI. Acknowledgements

We acknowledge Casey Stone, Austin Clyde, Defne Gorgun, Alexander Brace, Maxim Zvyagin, Ozan Gokdemir, Kiayi Shao, Ashka Shah, Rory Butler, Rebecca Weinberg, Carla Mann, Heng Ma, Abraham Stroka, Gyorgy Babnigg, Dion Antonopolous, Thomas Brettin, James Davis, and Christopher Fry for helpful discussions and implementing the automated lab/infrastructure. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the US Government.

REFERENCES

[1] R. Stevens, V. Taylor, J.A. Nichols, A.B. MacCabe, K. Yelick, and D. Brown. Ai for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science. Technical report, Department of Energy, 2019.

[2] Priya Choudhry. High-throughput method for automated colony and cell counting by digital image analysis based on edge detection. *PLOS ONE*, 11(2):1–23, 02 2016.

[3] Yu-Heng Cheng, Yu-Chih Chen, Riley Brien, and Euisik Yoon. Scaling and automation of a high-throughput single-cell-derived tumor sphere assay chip. *Lab Chip*, 16:3708–3717, 2016.

[4] Ainhoa Letamendia, Celia Quevedo, Izaskun Ibarbia, Juan M. Virto, Olaia Holgado, Maria Diez, Juan Carlos Izpisua Belmonte, and Carles Callol-Massot. Development and validation of an automated high-throughput system for zebrafish in vivo screenings. *PLOS ONE*, 7(5):1–13, 05 2012.

[5] Jörn Glökler, Tatjana Schütze, and Zoltán Konthur. Automation in the high-throughput selection of random combinatorial libraries—different approaches for select applications. *Molecules*, 15(4):2478–2490, 2010.

[6] Lei Xiao. Designing and implementing a large-scale high-throughput total laboratory automation (tla) system for dna database construction. *Forensic Science International*, 302:109859, 2019.

[7] Liwei Cao, Danilo Russo, Kobi Felton, Daniel Salley, Abhishek Sharma, Graham Keenan, Werner Mauer, Huanhuan Gao, Leroy Cronin, and Alexei A. Lapkin. Optimization of formulations using robotic experiments driven by machine learning doe. *Cell Reports Physical Science*, 2(1):100295, 2021.

[8] Jarosław M. Granda, Liva Donina, Vincenza Dragone, De-Liang Long, and Leroy Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, 559(7714):377–381, 2018.

[9] Adarsh Dave, Jared Mitchell, Kirthevasan Kandasamy, Han Wang, Sven Burke, Biswajit Paria, Barnabás Póczos, Jay Whitacre, and Venkatasubramanian Viswanathan. Autonomous discovery of battery electrolytes with robotic experimentation and machine learning. *Cell Reports Physical Science*, 1(12):100264, 2020.

[10] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–89, 2009.

[11] Anthony Coutant, Katherine Roper, Daniel Trejo-Banos, Dominique Bouthinon, Martin Carpenter, Jacek Grzebyta, Guillaume Santini, Henry Soldano, Mohamed Elati, Jan Ramon, Celine Rouveirol, Larisa N. Soldatova, and Ross D. King. Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proceedings of the National Academy of Sciences*, 116(36):18142–18147, 2019.

[12] Kevin Williams, Elizabeth Bilsland, Andrew Sparkes, Wayne Aubrey, Michael Young, Larisa N. Soldatova, Kurt De Grave, Jan Ramon, Michaela de Clare, Worachart Sirawaraporn, Stephen G. Oliver, and Ross D. King. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of The Royal Society Interface*, 12(104):20141289, 2015.

[13] Ross D. King, Vlad Schuler Costa, Chris Mellingwood, and Larisa N. Soldatova. Automating sciences: Philosophical and social dimensions. *IEEE Technology and Society Magazine*, 37(1):40–46, 2018.

[14] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[16] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

[17] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[18] Sumit Kumar Jha, Arvind Ramanathan, Rickard Ewetz, Alvaro Velasquez, and Susmit Jha. Protein folding neural networks are not robust. *arXiv preprint arXiv:2109.04460*, 2021.

[19] Leonard T. Nguyen, Evan F. Haney, and Hans J. Vogel. The expanding scope of antimicrobial peptide structures and their modes of action. *Trends in Biotechnology*, 29(9):464–472, 2011.

[20] Payel Das, Tom Sercu, Kahini Wadhawan, Inkit Padhi, Sebastian Gehrmann, Flaviu Cipcigan, Vijil Chenthamarakshan, Hendrik Strobelt, Cicero dos Santos, Pin-Yu Chen, Yi Yan Yang, Jeremy P. K. Tan, James Hedrick, Jason Crain, and Aleksandra Mojsilovic. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, 5(6):613–623, 2021.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Scott N. Dean and Scott A. Walper. Variational autoencoder for generation of antimicrobial peptides. *ACS Omega*, 5(33):20746–20754, 08 2020.

[23] Zheng Gong, Yanze Wu, Liang Wu, and Huai Sun. Predicting thermodynamic properties of alkanes by high-throughput force field simulation and machine learning. *Journal of Chemical Information and Modeling*, 58(12):2502–2516, 2018. PMID: 30205676.

[24] Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*, 2020.

[25] George Gondim-Ribeiro, Pedro Tabacof, and Eduardo Valle. Adversarial attacks on variational autoencoders. *arXiv preprint arXiv:1806.04646*, 2018.

[26] Meysam Sadeghi and Erik G Larsson. Physical adversarial attacks against end-to-end autoencoder communication systems. *IEEE Communications Letters*, 23(5):847–850, 2019.

[27] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

[28] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[29] Arvind Ramanathan, Laura L Pullum, Faraz Hussain, Dwaipayan Chakrabarty, and Sumit Kumar Jha. Integrating symbolic and statistical methods for testing intelligent systems: Applications to machine learning and computer vision. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 786–791. IEEE, 2016.

[30] Sumit Jha, Rickard Ewetz, Alvaro Velasquez, and Susmit Jha. On smoother attributions using neural stochastic differential equations. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 522–528. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[31] Jianli Pan and Zhicheng Yang. Cybersecurity challenges and opportunities in the new "edge computing + iot" world. In *Proceedings of the 2018 ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization*, SDN-NFV Sec'18, page 29–32, New York, NY, USA, 2018. Association for Computing Machinery.

[32] Yaakov Katz and W Jpost. Stuxnet virus set back iran's nuclear program by 2 years. *Jerusalem Post*, 15, 2010.

[33] L Nonboe Andersen, Jan Larsen, L Kai Hansen, and Mads Hintz-Madsen. Adaptive regularization of neural classifiers. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 24–33. IEEE, 1997.

[34] Jan Larsen, L Nonboe, Mads Hintz-Madsen, and Lars Kai Hansen. Design of robust neural network classifiers. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 1205–1208. IEEE, 1998.

[35] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017.

[36] Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Oleg Sokolsky, and Insup Lee. Detecting oods as datapoints with high uncertainty. *arXiv preprint arXiv:2108.06380*, 2021.