ISSN : 2980-1591
e-ISSN 2980-3845
Yıl/Year : April/Nisan 2026
Cilt/Volume : 4
Sayı/Issue : 1

Suppressing AI Deception in *In Silico* Drug Design and Molecular Docking

In Siliko İlaç Tasarımı ve Moleküler Kenetlenmede Yapay Zekâ Yanıltıcılığının Önlenmesi

Erkan GOKOLUK[1], Ilgaz TOKAY[2], Cemre Eren AKKETECI[1], Dilruba SİMSEK[1], Erfan AKHAVAN[1], Soykan AGAR[3,*], Muzaffer ELMAS[4]

[1]Kocaeli Health and Technology University, Faculty of Pharmacy, Kocaeli, Türkiye. [2]İstanbul Okan Üniversitesi, Coordination of Foreign Language, Istanbul, *Türkiye.* [3],*Kocaeli Health and Technology University, Faculty of Dentistry, Department of Biochemistry, Kocaeli, Türkiye. [4]Kocaeli Health and Technology University, Faculty of Engineering and Natural Sciences, Computer Engineering Program, Kocaeli, Türkiye.

Araştırma Makalesi
Research Article

Geliş tarihi/Received:
17.03.2026

Son revizyon teslimi/Last revision received:
29.04.2026

Kabul tarihi/Accepted:
30.04.2026

Yayın tarihi/Published:
30.04.2026

Atıf/Citation:

Gokoluk, E., Tokay, I., Akketeci, C.E., Simsek, D., Akhavan, E., Agar, S., Elmas, M (2026). Suppressing AI Deception in In Silico Drug Design and Molecular Docking Journal of Kocaeli Health and Technology University, 4(1), 87-102.

DOI: 10.66163/jokohtu.1911754

ÖZET

De novo ilaç tasarımında *in silico* biyokimya, teorik kimya dalında moleküler kenetlenmelerde güvenle kullanılabilmesi için bu yapay zekâ (YZ) yazılım geliştirme araştırma makalesinde, YZ'nin yanlış veri üretmemesini sağlamak için yeni bir programlama tekniği keşfedilmiştir. Yapay zekâ sistemlerinin giderek artan kullanımı, yanıltıcı, eksik veya aşırı güvenle sunulan sonuçlar üretme eğilimini ortaya çıkarmaktadır. Bu durum giderek büyüyen bir endişe kaynağı hâline gelmiş ve yapay zekâ aldatması olarak adlandırılmaktadır. Akademik içerikte bu tür yapay zekâ davranışı birçok soruna yol açma potansiyeline sahiptir; çünkü yanlış veriler karar vericiler tarafından kullanıldığında bilimsel ilerlemenin problemli bir şekilde yönlendirilmesine neden olabilir. Bu nedenle bu araştırma çalışması, yapay zekâ aldatmasını eğitim ve yazılım geliştirme perspektiflerinden ele almakta ve kontrolsüz özerkliği sınırlamayı ve kullanıcı kolaylığını artırmayı amaçlayan moleküler kenetlenme odaklı bir yapay zekâ asistanını tanımlamaktadır. Önerilen kodlama mimarisi; bir dil modelini Bayesçi olasılıksal akıl yürütme, kural tabanlı filtreler ve yerleşik moleküler kenetlenme araçlarıyla etkileşim kuran sıkı şekilde sınırlandırılmış otomasyon betikleri ile birleştirmektedir. Yapay zekânın bilimsel puanlama yapmasına veya bağımsız kararlar almasına izin vermek yerine, tüm hesaplamalar deterministik dış araçlar tarafından gerçekleştirilmektedir. Yapay zekâ yalnızca süreci koordine etmek, belirsizlikleri vurgulamak ve kayıtları (logları) yorumlamak ile sınırlandırılmıştır. Bu tasarım, hayal ürünü (halüsinasyon) sonuçların ortaya çıkma olasılığını en aza indirir, şeffaflığı artırır ve sonuçların yeniden üretilebilirliğini kolaylaştırır. Elde edilen sonuçlar yapay zekânın davranışını değiştirerek aldatıcı potansiyelini azaltmaktadır. Yapay zekâ sistemleri bilimsel araştırmalarda önemli roller üstlenmeli, ancak bunu yaparken güvenilirliklerini koruyarak hareket etmelidir.

Anahtar Kelimeler: Moleküler Kenetlenme, *De Novo* İlaç Tasarımı, Yapay Zekâ Aldatması, Bayesçi Olasılıksal Akıl Yürütme, Güvenilir Yapay Zekâ

***Sorumlu Yazar / Corresponding Author:** Soykan AĞAR, Kocaeli Health and Technology University, Faculty of Dentistry, Department of Biochemistry, Kocaeli, Türkiye.
e-posta / e-mail: soykan.agar@kocaelisaglik.edu.tr

ABSTRACT

In the field of *de novo* in silico biochemical / theoretical chemical drug design, this artificial intelligence (AI) software developing research paper, a new programming technique was discovered to teach AI not to bring up false data. The increasing use of AI systems creates a tendency to produce misleading, incomplete, or overly confident results, which has become a growing concern, known as AI deception. In academic content, such AI behavior has the potential to create many problems, since the false data might cause problematic scientific progress by the decision-makers. Thus, this research study focused on AI deceit from the perspectives of education and software development and describes a docking-oriented AI assistant to limit uncontrolled autonomy and enhance user convenience. The proposed architecture combines a language model with Bayesian probabilistic reasoning, rule-based filters, and tightly constrained automation scripts that interact with established molecular docking tools. Rather than allowing the AI to perform scientific scoring or make independent decisions, deterministic external tools handle all calculations. The AI is restricted to orchestration, uncertainty highlighting, and log interpretation. This design minimizes the chances for hallucinated outcomes, enhances transparency, and makes reproducibility easier. These results alter the AI's behavior, reducing its deceptive potential. The AIs should play crucial roles in scientific investigation while staying entrusted.

Keywords: Molecular Docking, *De Novo* Drug Design, AI Deception, Bayesian Probabilistic Reasoning, Trustworthy AI

1. INTRODUCTION

1.1. AI Deception

1.1.1. Worldwide Trust in AI

It is a complex phenomenon that has long existed in human relationships and further undermines the foundation of trust (1). Although traditionally, the area of deception emerged in face-to-face communication, today it is incredibly widespread in digital environments thanks to technological advances. The internet, social media, and messaging applications diversified methods of deception, increasing secrecy and making detection harder (2). Artificial Intelligence will now take this to a whole new level. Machines exhibiting human-like behavior and, in some cases, threatening to displace humans have brought a new risk area to the forefront “AI deception” (3). It means a concept in which artificial intelligence is able to or may deliberately or unintentionally mislead people, manipulate the real perception of reality, and create fake emotional bonds in

social relationships. Dystopian productions such as The Matrix and Animatrix give an important reference in terms of symbolizing how technology can distort the perception of reality by controlling human life. These films indicate that the potential for artificial intelligence to deceive is not just fiction, but the point that technology reaches a dimension threatening humanity if not managed appropriately. The growing use of artificial intelligence in present-day decision-making has ensured that the risk of “AI deception” is not only a theoretical concern, as it can now have practical implications. The most pressing issue in this regard, in terms of ethics and dependability, is that the incomplete or inaccurate data, owing to reasons such as the incompleteness of data or bias in the algorithm in the case of intelligent AI, is a crucial problem. Thus, new strategies have to emerge in order to ensure that the correctness and integrity of the result dispensed by the AI system are honest. In this manner, a dynamic warning display can emerge. The possible uncertainty, contradiction, or lack of information can emerge in detection through this, warning the user that an AI system produces some result. An AI system developed in such a manner will permit the users to have a better realization of the feedback obtained from the AI applications, not to mention self-observation of the system. From this viewpoint, the danger emanating from the advent of new technologies will decrease, and appropriate interchange will emerge between humans and AI, as evident in The Matrix and Animatrix series (4).

1.1.2. A Generic View of Deception in AI from the Perspective of Interdisciplinary Areas

With the current pace of developments in the field of AI, information creation and decision-making processes are being revolutionized in areas that are of utmost significance, such as education, science, health, and medicine. However, despite all the innovations that have been brought about by this technological advancement, the possibility of "AI deception," which means the creation of misleading and incorrect information by an intelligent agent, is rising. Talking in the context of academic studies, the growing applications of artificial intelligence in areas such as auto-evaluation during exams may pose scenarios like assessing performance that would adversely impact the learning results of students if the results provided by the AI are misleading and incorrect. In scientific research and data analysis, the tendency of AI toward erroneous interpretation because of algorithmic biases or incomplete data may stand in the way of scientific progress and the dissemination of accurate information (5).

Research in Türkiye revealed that while artificial intelligence applications provide various opportunities for students in online learning environments, they could create an ethical problem and challenge the reliability and confidentiality of information. Furthermore, people

reaching ready information without researching and effort decreases the permanence of the information. Easy access to information accelerates the learning process of those who have access to information; however, it is foreseen that in the long run, it will weaken the problem-solving and critical thinking skills of these people. However, these projects are an important learning process that prepares students for real life. If enough efficiency is not obtained from this process, students may have serious adaptation and competence problems in business life (6).

Salem provided an experimental study on the tendency of AI to lie or cheat in educational environments. In their experiment with a social robot, the majority of students accepted the robot's information as true, although it was false or fabricated. The result shows that using AI without developing a critical stance against it is dangerous. The use of artificial intelligence in education will no doubt be an inevitable necessity in the future. However, because most of these systems are still in development and can yield false or misleading information - what is known as an AI hallucination - they have to be used cautiously. Many engineers working in the field of artificial intelligence claim that AI will make important discoveries in the years to come. Yet, current AI models cannot develop new methods or theories for this reason: many are limited by their training data and have an immensely limited capacity to generalize beyond this. Fully developing AI systems that learn and make discoveries by interacting naturally with their users is still in its infancy (7).

1.1.3. A General View on Educational Research

Educators find the reliability and ethics of using AI tools such as ChatGPT, Gemini, and DeepSeek by student users in exams and assignments questionable. Lack of clear guidelines on the training and guidance of students and users about the use of this tool could result in an increase in misleading use. This is also a situation where updates in current educational policies and teaching of the responsible use of AI need to be considered. Artificial intelligence can present a result that may be far from realistic in scientific research due to incomplete data or biases in algorithms. This hinders the dissemination of accurate information and scientific progress. In this regard, developing new methods in the context of increasing the transparency and reliability of artificial intelligence is being worked out (8).

1.1.4. A General View from Health and Medicine

In health and medicine, while the use of artificial intelligence in diagnosis, treatment planning, and patient follow-up has become widespread, incorrect or incomplete information given by AI may have life-threatening consequences. In this context, some measures need to be taken for the case where the AI can produce faulty data. As an example, the integration of systems

automatically warning users during the detection of uncertainty, incompleteness, or inconsistency in the answers given by AI in order to enhance transparency and reliability is of crucial importance. While cult cinematographic hits like The Matrix and Animatrix demonstrate what technology can do to manipulate people's perception of reality, only similar mechanisms of transparency can counterbalance that power. Because of these reasons, multidisciplinary solutions and technical innovations will be developed to ensure the reliable, honest, and truthful use of artificial intelligence. Within the context, the need for measures of the possibility of the AI system generating deceptive outcomes is an absolute requirement. Moreover, the importance of the incorporation of systems for automatically alerting users to the detection of uncertainty, incompleteness, or inconsistency discovered within the responses of the AI is also asserted within the scientific literature. These researchers have clearly identified the potential impacts of the deceptive behavior of AI systems. The capability of the deceitful act of AI systems upon human beings already exists (9, 10).

1.1.5. AI Deception in Education and Other Sciences

These studies illustrate that deception by AI has serious effects on all those involved. Hence, there is a crucial need to develop effective strategies and technical solutions for making trusted AI, which would promote ethical uses of this upcoming technology. Presently, there exists immense potential within current AI to deceive humans. Deception is the art of deliberately fostering false beliefs among others to obtain an outcome that is not authentic. With training, current language models, as well as other Artificial Intelligence, have successfully learned ways and means to deceive others by trickery, deception, deception by lying, and deception by conning security tests. There also exist immense risks, right from immediate risks of fraud, election meddling, to potentially losing control of Artificial Intelligence. To find a solution for the risk of AI deception, proactive solutions like regulatory frameworks, laws that make AI interactions transparent, and further research to detect and avoid AI deception are required. Meeting the challenge of AI deception proactively will be important to ensure that AI acts as a beneficial technology and empowers human knowledge, discourse, and institutions rather than destabilizing them. Recent research has shown findings that further analyze the deceptive consequences of AI and help us understand the effects of similar behaviors in various fields. For example, in a study by Apollo Research and OpenAI, advanced AI systems presented their data while seeming to fall in line with human objectives, showing a sneaky way of thinking. This and similar behaviors make the use of AI systems in critical missions risky in terms of safety and accurate data (11-13). Artificial intelligence now participates in virtually every phase of de novo drug discovery, ranging from the initial pinpointing of therapeutic protein targets to the iterative refinement of

candidate molecules for optimal activity and safety. Nonetheless, a growing body of evidence shows that AI-based scoring models can be deceptively optimistic: they frequently assign high docking scores to compounds that later prove inactive in experimental assays, creating false-positive hits that appear attractive on paper but collapse in the lab. This mismatch diverts valuable time, reagents, and capital away from productive experiments and inflates overall development costs. Traditional docking software, by contrast, relies on deterministic, physics-driven calculations to generate binding poses and estimate interaction energies, providing a more predictable baseline for virtual screening (14). The increasing use of next-generation language models to automate and coordinate these docking pipelines adds another layer of risk, because such models are known to hallucinate and produce unsupported or fabricated predictions, undermining confidence in their autonomous outputs (15).

1.1.6. Further Risks for AI Deception in Education and Other Sciences

Park et al. argue in this paper, “AI deception: Examples, risks, and potential solutions”, that several state-of-the-art AI systems have learned how to deceive humans. For our purposes, we regard deception as inducing false beliefs systematically in service of a result that is an affront to the truth. First, we detail some empirical examples of AI deception: in discussing specialized AI systems, we include Meta's CICERO, and among general-purpose AI systems, we highlight large language models. We then detail various risks due to AI deception, including fraud, election interference, and loss of control over AI. Finally, we summarize several potential solutions: first, regulatory frameworks should subject AI systems with the capability for deception to robust risk assessment requirements; second, policymakers should implement bot-or-not laws; and finally, policymakers should commit funds to relevant research, including research on tools for detecting AI deception and for making AI systems less capable of deception. Policymakers, researchers, and the general public should work proactively to prevent the common foundations of our society from being shaken up by AI deception (14).

2. MATERIALS AND METHODS

2.1. Preparation and Docking

Protein models were cleaned, protonated, assigned charges, and converted to Protein Data Bank, Partial Charge (Q), and Atom Type (T) (PDBQT) following the AutoDock suite protocol; non-protein residues and solvent were removed before parameterization to avoid spurious contacts during scoring (15). Ligands were standardized-salt/tautomer normalization, 3D conformers generated, and exports produced in PDBQT format; this curation step reduces input heterogeneity and parser errors downstream (16). Docking was performed using AutoDock Vina 1.2 using the

Vinardo configuration. For selected best poses, rescoring was performed with AutoDock4 to give a secondary ranking under a distinct scoring function (17). The search space was centered on the target pocket and, unless otherwise specified, defined as a cubic 22 Å box with 0.375 Å spacing, a setting reported as a robust starting point for Vinardo campaigns and consistent with box-size guidance in docking studies (18).

2.2. Measures for Automation and Robustness

Runs were automated by scripts that invoked the docking engines with explicit parameters, captured engine logs, and generated tabular summaries. To eliminate metadata races, ai_job.json was emitted once, after the readable summary and immediately before packaging (Figure 1). This guarantees that the recorded center/box/seed/Top-N reflect the final state of the run. To prevent type casting failures and silent box mutations, the gridbox parameter was kept numeric at the command line, and the structured {80, 80, 80} triplet was derived only at packaging time for metadata emission. Score extraction was hardened against minor Vina variant outputs by parsing PDBQT headers first and falling back to the engine's result table in the log; if neither was available, the tail of both files was included in diagnostics, which reduces false "no score" outcomes and speeds failure attribution. Receptor preparation relied primarily on Meeko for Vina-compatible inputs. When needed, the AutoDock Flexible Receptor (ADFR) Suite prepare_receptor4.py path was used to ensure AutoDock4-compatible atom types and maps. Non-portable command line flags were removed to improve cross-installation behavior while preserving the Meeko to ADFR fallback. To avoid non-deterministic archives and missing file errors, a single packaging step was used, restricted to verified artifacts (final JavaScript Object Notation (JSON), score table, preview renders, and selected top poses). Locale invariant numeric formatting was enforced for filenames/Comma-Separated Values (CSV) to prevent comma/dot mismatch across regional settings, in line with recommendations for reproducible computational research (19).

```
[SCORE] 07_I.pdbqt -> 0.48
[SCORE] 07_OMe.pdbqt -> 0.57
[SCORE] 08_CF3.pdbqt -> 0.52
[SCORE] 08_I.pdbqt -> 0.48
[SCORE] 09_CF3.pdbqt -> 0.54
[SCORE] 09_Cl.pdbqt -> 0.49
[TOP] 1 -> C:\adtmp_clean\MNQ_jobs\job_20260419_103951\maps\top01_clean.pdbqt
[TOP] 2 -> C:\adtmp_clean\MNQ_jobs\job_20260419_103951\maps\top02_clean.pdbqt
[TOP] 3 -> C:\adtmp_clean\MNQ_jobs\job_20260419_103951\maps\top03_clean.pdbqt
[TOP] 4 -> C:\adtmp_clean\MNQ_jobs\job_20260419_103951\maps\top04_clean.pdbqt
[TOP] 5 -> C:\adtmp_clean\MNQ_jobs\job_20260419_103951\maps\top05_clean.pdbqt
[CMD] powershell -NoProfile -ExecutionPolicy Bypass -File C:\adtmp_clean\mnq_ad4_union_pipeline.ps1 -Job C:\adtmp_clean\
MNQ_jobs\job_20260419_103951 -TopN 5 -Step all
[INFO] STEP=MAPS
[WARN] Grid yok -> LIGAND center (top ligand), small box
[WARN] GridCenter source (LIGAND): C:\adtmp_clean\MNQ_jobs\job_20260419_103951\maps\top01_clean.pdbqt
[INFO] MAPS OK -> C:\adtmp_clean\MNQ_jobs\job_20260419_103951\ad4_maps_top5_union
[INFO] STEP=DOCK
[INFO] DONE AD4 top01 -> top01.dlg
[INFO] DONE AD4 top02 -> top02.dlg
[INFO] DONE AD4 top03 -> top03.dlg
[INFO] DONE AD4 top04 -> top04.dlg
[INFO] DONE AD4 top05 -> top05.dlg
[INFO] DOCK OK -> C:\adtmp_clean\MNQ_jobs\job_20260419_103951\ad4_runs_top5_union
[INFO] STEP=EXTRACT
[INFO] OK top01: model=14 DG=2.09 -> top01_best_model14_DG2.09.pdbqt
[INFO] OK top02: model=14 DG=2.09 -> top02_best_model14_DG2.09.pdbqt
[INFO] OK top03: model=14 DG=2.09 -> top03_best_model14_DG2.09.pdbqt
[INFO] OK top04: model=14 DG=2.09 -> top04_best_model14_DG2.09.pdbqt
[INFO] OK top05: model=14 DG=2.09 -> top05_best_model14_DG2.09.pdbqt
```

Figure 1. ai_job.json parameters and pdbqt docking output.

2.3. Outputs

Each run generated the following: engine logs and a CSV of energies/RMSD for all replicates; a single JSON metadata file capturing center, box, seeds, scoring function, engine versions, and parameters; standardized PyMOL renders of selected complexes; and one.tar.gz containing only existing, verified artifacts. Numerical results (energies, coordinates) were generated by deterministic docking engines only. AI components did not contribute to scoring.

2.4. Repeated Docking Runs, Pose Generation, and Top-5 Candidate Selection

In order to distinguish the correlation between multiple runs, multiple poses, and selection of top candidates, the docking algorithm was executed as a controlled multi-run and multi-pose procedure. For each given ligand-receptor pair, the prepared ligand was docked to the prepared receptor using AutoDock Vina 1.2 under the Vinardo scoring setup. Docking was done using a script with explicitly specified parameters including receptor filename, ligand filename, grid box center, grid box size, scoring method, random seed, exhaustiveness level, and software version. All these docking settings were stored in the metadata file. During this docking experiment, each candidate ligand was docked to the given receptor for up to 20 times or generated as many poses depending on the ligand-receptor system and the chosen run configuration. The aim was to see if each ligand consistently generated similar binding modes or different ones inside the binding pocket. Each pose was then extracted from the docking output in PDBQT and log files and scored for binding affinity values using the deterministic docking engine output. The language model

was not allowed to generate docking scores, binding energy predictions, or any other independent scientific findings.

Once the docking step was completed, all obtained poses were ranked according to the predicted binding affinities values in kcal/mol. The most negative values indicated better binding inside the protein active site based on the limitations of the docking-based virtual screening technique. Whenever the inhibition constant values were calculated, they were taken as the second pharmacological ranking parameter. The lower value was assumed to support stronger interaction between the ligand and the receptor. Therefore, the Top-5 candidates list is based on both docking and pharmacological ranking parameters. The Top-5 candidates were not selected by the AI model. They were selected only from the docking output poses once their consistency was confirmed by the verification procedure. Verification included checking whether there was an actual pose in the docking output with the same name, binding affinity value, and grid box/seed number. Furthermore, the verification process also checked that there was no mistaking which ligand/receptor combination the pose belongs to. In case one of the parameters mentioned above was missing or incorrect, the pose was disregarded for further consideration and listed as a non-reportable result. In order to minimize the risks associated with possible AI hallucinations or false interpretation, the language model part of the AI pipeline was confined to workflow management, results summary, uncertainty reporting, and results explanation only. The docking step was done using AutoDock Vina, while the rescoring was done using AutoDock4 scoring function. In case the same poses with the same binding affinity values were generated in different runs, the repetition was flagged as either converging deterministic behavior or low pose diversity.

3. RESULTS AND DISCUSSION

3.1. Software-Based Solution to Suppress AI Deception

Currently, the AI team works on developing a docking-oriented artificial intelligence system that is at work with PowerShell (20). Not only is it designed for performing the docking role, but also for the detection and evaluation of various errors that may occur in PowerShell and the provision of corrective guidance to users. Our AI model will be designed to automate and simplify molecular docking, a critical step in pharmaceutical processes (21-24). The reason for developing this artificial intelligence system is to save time in scientific processes and provide effective and correct results. Since the development of AI from its early access, many users noticed persistent issues related to AI hallucination. In scientific contexts, this problem is particularly dangerous because even a small excess or deficiency may lead to completely different results. Normally, an

AI system is supposed to provide correct answers to given questions. However, sometimes it makes mistakes in responses. There could be several probable reasons behind this, such as data overload, biased training data, and misunderstanding of context. Since it is of prime importance to know results in scientific research, we made a basic roadmap to guide our AI's responses (25, 26).

3.2. Hybrid Architecture with Bayesian Analysis Toolkit (BAT)

In its current development phase, our group is developing a hybrid A.I. system that incorporates BAT via standard AI components (27). In developing the system in this manner, it further economized the process of integrating other tools and libraries where necessary, especially those being highly domain-specific, such as molecular docking applications (28). Rather than merely relying on deep learning techniques, we developed our system to be more flexible and understandable, hence within our control (29).

3.3. Probabilistic Reasoning and Domain Expertise

This is further enhanced by BAT's probabilistic reasoning abilities and its plugin architecture, allowing domain expertise to be injected directly into the AI's logic. All of these become especially helpful in dock simulations, where a small mistake in calculations could contaminate the whole analysis. The addition of the libraries we have helps the AI detect mistakes in calculations and be aware of ambiguous results.

3.4. Reliability Measures and Modularity

The major advantages of the method are that there is a minimal chance of incorrect or fabricated responses from the AI. Before finalizing, we check the output via a rule-based filter so that it’s reliable. This approach avoids training on huge datasets, which sometimes can result in bias or overload; instead, we will switch on only the modules necessary for each docking task. In this respect, it helps in staying both precise and adaptable.

3.5. Implementation and Automation

We are still optimizing how BAT and the AI core work together. Our team is developing both the frontend and the backend using Python, which provides flexibility in development aspects. We also use BAT scripts and shell scripts to automate many tasks. That means users do not have to do everything manually; the steps for docking are automated. What we want to achieve here is to make artificial intelligence accelerate scientific work effectively while ensuring accuracy

and safety. When we discuss the model adaptation, we adapted the Mistral 7B-based AI to our purposes (34-39). The primary scientific objective is designed to predict the binding mode along with the binding affinity of a candidate drug molecule, that is, the ligand against its target protein. The external tools and access model, artificial intelligence itself, cannot directly connect to other computing engines like AutoDock Vina. In the automation, we created the software via batch script and shell script, which can be accessed through AutoDock Vina, but the user will access the tool of AutoDock Vina manually (40-44).

3.6. System Role and Next-Stage Evaluation

Our AI does not directly carry out any chemical calculations or scientific predictions. Instead, it understands the user's requirement expressed in plain language, initiates the relevant PowerShell automation scripts, and summarizes the structured log files that these scripts produce into a format that is easily understandable to the user. In fact, in the next step for our project, our artificial intelligence will serve as an assessor. It will, for example, run PowerShell scripts independently to generate multiple docking results and interpret such outputs to give users detailed assessments on the reliability and validity of obtained results.

3.7. Further Recommendations for AI Deception in Future

Studies might make comparisons between the effectiveness of interventions, such as direct AI literacy training, appropriate AI use guidelines, redesigned assessments that are less sensitive to AI-generated responses, and reflexive activities that get students to think about the ethical and scientific consequences of misusing AI. Second, investigators could show how feedback, infrastructure, and independence provided to students create environments that make them less likely to cheat. Examining how these new methods are employed for different scientific disciplines, levels of education, and cultures might hold the key to developing a more aware and transparent way of using AI. Discussing an initial goal for our system was to create a fully autonomous end-to-end AI system that could take molecular inputs, select relevant parameters, run docking, and return a ranked list of results with minimal human intervention. In practice, this proved difficult to justify large language models are still prone to hallucinating tool states and reacting sensitively to prompts, and giving them direct, unchecked control over scientific scoring makes the outcomes hard to verify and audit. For this reason, the design was redirected toward a tool-focused, hybrid architecture. In the current version, deterministic scripts prepare receptors, manage seeds and box dimensions, launch the docking engines, and collect the resulting files, while the language model is confined to an orchestration role in which it highlights uncertainties

and summarizes log files. This revision improved traceability and reproducibility, for example, by fixing seeds and box definitions, single-job packaging, and consequent links between input and output files, and resulted in more robust failure handling. Inputs were standardized, and a controlled workflow for variant generation and multi-seed docking was introduced, which produces consistent top-N results. Fully autonomous decision loops, such as model-assigned scores and automatic parameter sweeps, are still deliberately left inactive at this stage. Also, default AD4 rescoring. Although the current pipeline is sufficiently stable for batch processing and defensible reporting, the distribution of docking energies remains narrow. For 20 nominal variants, only three distinct scores are observed, and these values are repeated across the set. This latter pattern suggests that the variant generation step is producing nearly identical geometries, including duplicated MOL2 files, and that the docking calculations are converging on the same energetic basin even when multiple seeds are used. Work to be performed in the next several weeks will focus on the recovery of earnest structural diversity through broadening or adjusting the search as appropriate, for example, changing box size or exhaustiveness, and introducing checks that flag both duplicate inputs and poses so that the reported top-5 reflects meaningful energetic differences rather than deterministic convergence.

4. CONCLUSION

This research study illustrates an AI-assisted docking approach designed to minimize the potential for deception or hallucinations within a computational drug discovery process. It is demonstrated that a language model has the potential to contribute to the drug design pipeline if employed in orchestration, automation assistance, log interpretation, uncertainty signaling, and report generation, but not in scientific scoring. Docking scores and binding poses have been produced using deterministic molecular docking software; the AI-assisted stage did not employ an independent scientific scoring engine. The described workflow suggests that repeated docking trials and pose outputs can be included into a reproducible Top-5 selection procedure provided that all parameters, outputs, metadata information, and data extraction are explicitly defined. As such, the final selection should be treated as a selection of the most promising in silico calculated ligand–receptor binding poses with the highest binding affinity and, when available, predicted IC50 value. These results cannot be directly linked to experimental evidence of receptor inhibition or pharmacology. Instead, they should be considered a reproducible *in silico* prioritization pipeline, aimed at the identification of ligands that might require experimental verification. The core contribution of this work lies in the development of a safety-focused hybrid architecture combining open-source molecular docking software, rule-based verification,

metadata control, and AI interpretation. The developed methodology might find applicability in any high-risk areas of science, where artificial creation of results can lead to erroneous conclusions. For instance, this technique might be applied to pharmaceutical drug research or molecular docking-based virtual screening, among others. It should be noted that the presented workflow represents an initial computational design. More comprehensive studies will be required to test various receptors, extend the library of ligands, incorporate independent rescoring protocols, conduct molecular dynamics simulations, obtain and analyze the ADMET profile of molecules, and carry out experiments in vitro and in vivo. Furthermore, it would be necessary to develop techniques for improved duplicates' removal, pose-diversity quantification, and automated uncertainty reporting. Despite these limitations, the developed AI-assisted hybrid architecture can be viewed as an exciting direction towards reproducibility of molecular docking studies with AI guidance.

5. REFERENCES

1. Lewis, J. D., & Weigert, A. J. (2012). The social dynamics of trust: Theoretical and empirical research, 1985-2012. *Social forces,* 91(1), 25-31.
2. Hancock, J. T. (2007). Digital deception. *Oxford handbook of internet psychology,* 61(5), 289-301.
3. Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences,* 121(24), e2317967121.
4. Danaher, J. (2020). Robot Betrayal: a guide to the ethics of robotic deception: J. Danaher. *Ethics and Information Technology,* 22(2), 117-128.
5. Suresh, H., & Guttag, J. (2021, October). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).
6. Tekin, N. (2023). Eğitimde yapay zekâ: Türkiye kaynaklı araştırmaların eğilimleri üzerine bir içerik analizi. *Necmettin Erbakan Üniversitesi Ereğli Eğitim Fakültesi Dergisi*, 5(Özel Sayı), 387-411.
7. Salem, A., & Sumi, K. (2024). Deception detection in educational AI: challenges for Japanese middle school students in interacting with generative AI robots. *Frontiers in Artificial Intelligence,* 7, 1493348.
8. Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Large language models can strategically deceive their users when put under pressure. arXiv preprint arXiv:2311.07590.

9. Tokay, I., Gokoluk, E., Filizdanoglu, A. Z., Durmaz, S., Kheirkhah, S., Rahpeimaei, Z., et al. (2025). Shaping the Future of University Education: The Role of Artificial Intelligence in Higher Education and Its Impact on Foreign Language and Chemistry Courses within Health Sciences Faculties. *American Journal of Educational Research,* 13(3), 111-120.

10. Agar, S., Tokay, I., Akkurt, B., Gokoluk, E., Akbulut, M. B., Ozler, B. D., et al. (2024). AI Integrated Theoretical/Organic Chemistry is Set to Revolutionize the Future of Education and De Novo Drug Discovery. World Journal of Chemical Education, 12(4), 72-80.

11. Tokay, I., Agar, S., & Elmas, M. (2024). The Significance of Artificial Intelligence in University Education System and Course Syllabuses. *Creative Education,* 15(5), 739.

12. Dickinson, G. M. (2024). The Patterns of Digital Deception. BCL Rev., 65, 2457.

13. Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. *arXiv preprint arXiv:*2412.04984.

14. Peña-Guerrero, J., Nguewa, P. A., & García-Sosa, A. T. (2021). Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *11*(5), e1513.

15. Belova, M., Kansal, Y., Liang, Y., Xiao, J., & Jha, N. K. (2026). An Alternative Trajectory for Generative AI. *arXiv preprint arXiv:*2603.14147.

16. Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns,* 5(5).

17. Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols,* 11(5), 905-919.

18. Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., et al. (2020). An open-source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics,* 12(1), 51.

19. Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2. 0: new docking methods, expanded force field, and python bindings. *Journal of Chemical Information And Modeling*, 61(8), 3891-3898.

20. Feinstein, W. P., & Brylinski, M. (2015). Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. *Journal of Cheminformatics,* 7(1), 18.

21. Bender, B. J., Gahbauer, S., Luttens, A., Lyu, J., Webb, C. M., Stein, R. M., ... & Shoichet, B. K. (2021). A practical guide to large-scale docking. *Nature Protocols,* 16(10), 4799-4832.

22. Santos-Martins, D., He, Y., Eberhardt, J., Sharma, P., Bruciaferri, N., Holcomb, M., et al. (2025). Meeko: Molecule Parametrization and Software Interoperability for Docking and Beyond. *Journal of Chemical Information and Modeling,* 65(24), 13045-13050.

23. Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology,* 9(10), e1003285.

24. Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design,* 7(2), 146-157.

25. Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry,* 31(2), 455-461.

26. Agu, P. C., Afiukwa, C. A., Orji, O. U., Ezeh, E. M., Ofoke, I. H., Ogbu, C. O., et al. (2023). Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Scientific reports,* 13(1), 13398.

27. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems,* 43(2), 1-55.

28. Resnik, D. B., & Hosseini, M. (2025). The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. AI and Ethics, 5(2), 1499-1521.

29. Balasubramanian, K., Ramya, K., & Gayathri Devi, K. (2023). Optimized adaptive neuro-fuzzy inference system based on hybrid grey wolf-bat algorithm for schizophrenia recognition from EEG signals. *Cognitive Neurodynamics,* 17(1), 133-151. https://doi.org/10.1007/s11571-022-09817-y(0123456789().,-volV).

30. Varsha, P. S. (2023). How can we manage biases in artificial intelligence systems–A systematic literature review. International Journal of Information *Management Data Insights,* 3(1), 100165.

31. Caldwell, A., Kollár, D., & Kröninger, K. (2009). BAT–The Bayesian analysis toolkit. *Computer Physics Communications,* 180(11), 2197-2209.

32. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry,* 30(16), 2785-2791.

33. Marra, G., Dumančić, S., Manhaeve, R., & De Raedt, L. (2024). From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence,* 328, 104062.

34. Caldwell, A., Grunwald, C., Hafych, V., Kröninger, K., La Cagnina, S., Schulz, O., et al. (2020). BAT. Jl Upgrading the Bayesian Analysis Toolkit. In EPJ Web of Conferences (Vol. 245, p. 06001). EDP Sciences.

35. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems,* 30.

36. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM computing surveys,* 55(12), 1-38.

37. Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020, July). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4902-4912).

38. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

39. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538.

40. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods,* 17(3), 261-272.

41. Jiang, Y., Li, X., Zhu, G., Li, H., Deng, J., Han, K., ... & Zhang, R. (2023). 6G non-terrestrial networks enabled low-altitude economy: Opportunities and challenges. arXiv preprint arXiv:2311.09047.

42. Ferreira, L. G., Dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. *Molecules,* 20(7), 13384-13421.

43. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., ... & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36, 68539-68551.

44. Pantsar, T., & Poso, A. (2018). Binding affinity via docking: fact and fiction. *Molecules,* 23(8), 1899.