Ensemble Multi-Modal Agentic AI Pipeline for Parkinson's Disease Drug Discovery

Anonymous Author(s)
Affiliation
Address
email

Abstract

Parkinson's disease (PD) remains without a cure, but recent advances in multi-modal AI offer new avenues for discovering disease-modifying treatments. We present a novel ensemble AI pipeline that integrates multiple state-of-the-art AI platforms – to identify and evaluate drug candidates for PD. Our system combines each platform's outputs using ensemble learning to overcome the limitations of any single model. Focusing on the hypothesis of enhancing glucocerebrosidase (GCase) activity , the pipeline discovered five novel small molecules. Each candidate was evaluated across mechanism of action, blood-brain barrier permeability, ADMET properties, toxicity, manufacturability, and patent novelty.These results underscore the potential of combining multi-modal foundation models and LLM agents to accelerate drug discovery

1 Introduction

Parkinson's disease is a neurodegenerative disorder characterized by dopaminergic neuron loss and pathological protein aggregates. No treatment to date can slow or stop PD progression, so discovering disease-modifying drugs remains a critical challenge. Mounting evidence links PD to lysosomal dysfunction: mutations in the GBA gene (encoding GCase) impair cellular waste disposal, leading to -synuclein buildup (Mullin et al., 2020). Enhancing GCase activity is therefore a promising therapeutic strategy.

Recent AI breakthroughs are transforming Parkinson's disease (PD) drug discovery. AlphaFold has unlocked accurate 3D structures of key PD proteins, accelerating structure-based design. Generative platforms like NVIDIA's BioNeMo and Google's TxGemma enable large-scale molecular design and analysis, while initiatives such as FutureHouse and Biomni deploy autonomous agents for automated discovery. Notably, companies like Insilico Medicine have already advanced AI-designed brain-penetrant compounds into preclinical testing. These advanced models offer unprecedented capabilities; however, a single AI model's predictions can be **unreliable or biased** if used in isolation. This highlights the need for **ensemble approaches** where multiple AI agents collaborate and cross-check each other's results.

In this work, we propose an **ensemble multi-modal AI pipeline** for drug discovery targeting PD. Our pipeline integrates four cutting-edge AI platforms: **TxGemma** for multi-modal understanding and property prediction, **FutureHouse** agents (Crow, Falcon, Owl, Phoenix) for literature mining, reasoning and experiment planning, and **Stanford Biomni** for orchestrating analysis tasks with its 150+ tools and databases. We combine their outputs using an ensemble learning strategy to identify promising drug candidates and filter them through a comprehensive set of criteria. **Results** highlight how our pipeline's synergy of LLM reasoning and predictive modeling led to promising PD therapeutic candidates.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

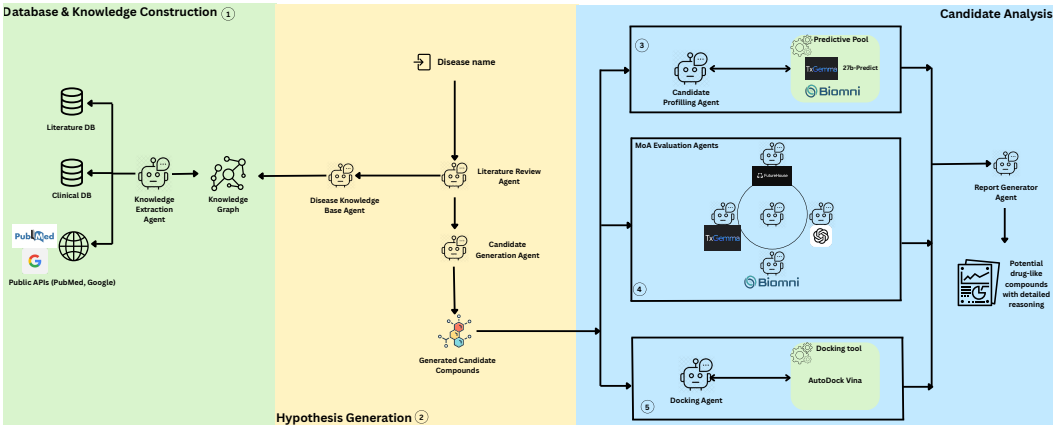

Figure 1: Ensemble multi-modal agentic pipeline

2 Methods

Our pipeline (Figure 1) comprises sequential AI-driven modules: **(1) Knowledge Extraction & MoA Modeling, (2) Generative Chemistry, (3) ADMET Filtering, (4) Mechanism-of-Action Screening,** and **(5) Docking-based Validation**. The process is orchestrated by an ensemble of agents that share information via a centralized workflow. We leveraged both structured predictive models and unstructured reasoning via LLMs, allowing decisions to be informed by numeric estimations as well as textual biomedical knowledge.

We began by deploying AI agents to extract biomedical data from literature, clinical reports, and databases, constructing a **knowledge graph (KG) of Parkinson's disease mechanisms and therapeutic strategies**. This KG allowed us to systematically identify promising hypotheses and mechanisms of action (MoA). Among the most compelling were **GCase chaperoning** and **lysosomal/autophagy activation**, both strongly linked to reducing -synuclein accumulation. These insights shaped the downstream stages of our pipeline, guiding compound generation and mechanistic validation.The pipeline's AI chemist module then generates candidate molecules that might have the desirable MoAs. We performed high-throughput *in silico* screening of the generated molecules for **absorption, distribution, metabolism, excretion, and toxicity (ADMET)** profiles. Using an ensemble of chemistry predictor, including TxGemma, Biomni (incorporating neural ADMET models and LLM evaluations), each compound was evaluated on key criteria: **BBB permeability**, **oral absorption**, **Ames mutagenicity**, **hERG cardiotoxicity**, **hepatotoxicity (DILI)**, **carcinogenicity**, **skin sensitization**, and **acute toxicity**. Compounds had to pass predefined thresholds (e.g. **BBB-permeable** and **non-toxic**) to advance. The candidates were then subjected to a novel **LLM-based MoA screening**. We posed four critical questions to an ensemble of reasoning models (including different LLM prompts and knowledge sources) (e.g. *Does the molecule bind misfolded GCase and enhance its ER→lysosome trafficking?)*.

Each question was answered by multiple LLM agents and a "decision" agent consolidated the responses. Only molecules receiving affirmative consensus were retained. Finally, we evaluated the top candidates with **molecular docking** to PD-relevant targets, focusing on **glucocerebrosidase (GCase)**. We obtained the crystal structure of GCase (PDB ID: 3GXI, 2.9Å) and prepared it for docking by defining the active site region (centered near the catalytic E340/E235 residues). Using AutoDock Vina, each candidate was flexibly docked into GCase's active site. We performed exhaustive docking runs (multiple binding site hypotheses) and recorded the best binding energies and poses for each compound. Additionally, we assessed docking to secondary targets if relevant (e.g. TRPML1 or other lysosomal proteins involved in autophagy) to check off-target interactions. The docking results were then analyzed by the pipeline's analysis agent, which also calculated the key physicochemical properties of the lead compound for context (molecular weight, logP, polar surface area, etc.). The final selection was made by balancing **docking score**, **mechanistic plausibility**, and **predicted drug-likeness**. Our ensemble pipeline's decision agent weighted these factors to choose the top candidate for reporting.

3 Results

The ensemble AI pipeline successfully generated five promising drug candidates for Parkinson's disease. The system **generated 1,106 novel candidates** and winnowed them down via multi-step filtering. After **ADMET screening**, 64 candidates remained, all predicted to be CNS-permeable and generally non-toxic. The subsequent **LLM-based MoA evaluation** identified **5 candidates** that met the top mechanistic criteria (predicted to enhance GCase folding/trafficking, stimulate lysosomal/autophagy pathways, etc.). Among these, **Compound 1** was therefore selected as the lead (molecular structure can be shared upon request). Compound 1 has many similarities with ambroxol, a GCase-targeting drug currently in Phase 3 trials.l. In silico analyses showed that Compound 1 preserved ambroxol's core mechanisms. The LLM ensemble predicted it would act as a GCase pharmacological chaperone, stabilize misfolded enzyme, and promote lysosomal/autophagy activation, supporting clearance of -synuclein and other toxic aggregates. Docking confirmed strong binding to GCase (–6.5 kcal/mol), slightly better than ambroxol (–6.2 kcal/mol), with the di-brominated ring and amino-cyclohexanone moiety forming favorable hydrophobic and hydrogen-bonding interactions. Compound 1 also demonstrated a favorable ADMET profile (see Table 1). Toxicity screens raised no major concerns, with only a moderate ($\tilde{3}0\%$) clinical risk score, comparable to ambroxol's scaffold. Predicted metabolic stability (t½ $\tilde{7}$.7 h) and manageable CYP450 interactions further support its drug-like profile. Overall, Compound 1 emerges as an orally bioavailable, brain-penetrant, and low-toxicity analogue of ambroxol, satisfying key prerequisites for a PD therapeutic candidate. Experimental validation will be required, but computational evidence highlights its strong potential.

Table 1: Key properties of lead compound vs. ambroxol (*in silico* predictions).

Property	**Compound 1 (Lead)**	**Ambroxol (reference)**
Docking Affinity to GCase	-6.5 kcal/mol	-6.2 kcal/mol (est.)
Predicted BBB Permeation	92% (high)	$\sim$88% (high)
Predicted Oral Absorption (HIA)	92% (high)	$\sim$90% (high)
Predicted Bioavailability	75%	$\sim$70% (est.)
Molecular Weight	406.12 Da	378.11 Da (actual)
cLogP (lipophilicity)	2.76	2.65 (exp.)
TPSA (polar surface area)	75.3 Å^2	62.7 Å^2 (calc.)
H-bond Donors / Acceptors	3 / 4	2 / 3
Predicted Toxicity Alerts	None major; clinical risk $\sim$30% (moderate)	None major (known safe in trials)

Benchmarking overview (Appendix A). We compared ensemble predictions against curated ground truth for two PD-relevant compounds. *Ambroxol*: 18 properties evaluated (12 exact matches, 3 partial, 3 mismatches; accuracy $\approx 83\%$), with key misses on PPBR, CYP2D6, and hERG inhibition. *Rasagiline*: 14 properties (8 exact, 2 partial, 4 mismatches; accuracy $\approx 71\%$), with larger errors in oral bioavailability, P-gp substrate classification, half-life, and lipophilicity. Overall, the ensemble is highly reliable on ambroxol and shows more deviations on rasagiline; a systematic tendency to overestimate permeability and half-life is observed for some CNS drugs. For GCase–ambroxol, docking is directionally consistent with neutral-pH inhibition at the active site; to better align with ground truth we plan to incorporate pH-aware protonation, post-docking rescoring (e.g., MM/GBSA), and short MD refinement. Full per-endpoint tables, and confusion matrices are provided in Appendix A.

4 Conclusion

We developed an ensemble AI pipeline that integrates knowledge graphs, reasoning LLMs, generative chemistry, and simulation tools to accelerate drug discovery for Parkinson's disease. The pipeline successfully identified Compound 1, an ambroxol-inspired analogue that preserves ambroxol's core mechanisms while improving predicted GCase binding, brain penetration, and drug-likeness. By combining multiple AI agents, our approach reduces the biases of individual models and rigorously filters candidates through mechanistic, ADMET, and docking validation.These results highlight the potential of multi-modal AI to generate disease-modifying candidates more efficiently than traditional

pipelines. While our findings are in silico and require experimental validation, Compound 1 emerges as a compelling lead, with properties consistent with an orally bioavailable, brain-penetrant PD therapy. Moving forward, synthesis and wet-lab testing will be essential to confirm efficacy and safety, while iterative feedback will further refine the pipeline.

References

[1] S. Mullin *et al.* Ambroxol for the Treatment of Patients With Parkinson Disease With and Without Glucocerebrosidase Gene Mutations: A Nonrandomized, Noncontrolled Trial. *JAMA Neurology*, 77(4):427–434, 2020. https://pubmed.ncbi.nlm.nih.gov/31930374/.

[2] Y. Zheng *et al.* Large Language Models in Drug Discovery and Development: From Disease Mechanisms to Clinical Trials. *arXiv preprint* arXiv:2409.04481, 2024. https://arxiv.org/abs/2409.04481.

[Vina License] AutoDock Vina License. "Apache License 2.0." https://vina.scripps.edu/license/.

[RCSB Usage Policy] RCSB PDB Usage Policies. "PDB data files are available under CC0 1.0." https://www.rcsb.org/pages/usage-policy.

[5] wwPDB Announcement: "PDB Core Archives adopt Creative Commons CC0 1.0." 2021. https://www.rcsb.org/news/feature/611e8d97ef055f03d1f222c6.

[Gemma Terms] Google. Gemma Terms of Use. https://ai.google.dev/gemma/terms.

[TxGemma Terms] Google. "Health AI Developer Foundations Terms (TxGemma)." https://developers.google.com/health-ai-developer-foundations/terms.

[RCSB 3GXI] RCSB PDB entry 3GXI: Crystal structure of acid β-glucosidase (GCase) at pH 5.5. https://www.rcsb.org/structure/3gxi.

[9] FutureHouse. "FutureHouse Platform: Superintelligent AI Agents for Scientific Discovery." May 1, 2025. https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents.

[10] FutureHouse. "Terms & Conditions." https://www.futurehouse.org/terms-of-service.

[11] K. Huang *et al.* "Biomni: A General-Purpose Biomedical AI Agent." *bioRxiv*, 2025. https://www.biorxiv.org/content/10.1101/2025.05.30.656746v1.

[Biomni GitHub] SNAP–Stanford. "Biomni: A General-Purpose Biomedical AI Agent (GitHub repository)." License: Apache-2.0. https://github.com/snap-stanford/Biomni.

A Benchmark Results

A.1 Ambroxol — Ensemble vs. Ground Truth

Table 2: Comparison between ensemble prediction and ground truth of Ambroxol.

Property	Ensemble Prediction	Ground Truth	Match?	Notes
Oral Bioavailability	75.43%	$\approx 70 - 80\%$	Yes	Within expected range.
HIA (Human Intestinal Absorption)	93.63%	Yes	Yes	Correct classification.
P-gp Substrate	27.92%	Not a substrate	Partial	Low probability but not a clear classification; close to correct.
Caco2 Permeability	$-5.04\,\mathrm{cm\,s^{-1}}$	$\approx 45 \times 10^{-6}\,\mathrm{cm\,s^{-1}}$	No	Different scales used; ensemble underestimates permeability.
BBB Penetration	94.13%	Yes	Yes	Consistent with clinical studies.
Plasma Protein Binding (PPBR)	44.73%	$\approx 80 - 90\%$	No	Ensemble significantly underestimates PPBR.
CYP3A4 Substrate	20.41%	Substrate	Partial	Correct direction, but weaker interaction predicted.
CYP2D6 Substrate	84.63%	Not a substrate	No	False positive.
Clearance	$7.45\,\mathrm{mL\,min^{-1}\,kg^{-1}}$	$\approx 8.1\,\mathrm{mL\,min^{-1}\,kg^{-1}}$	Yes	Accurate prediction.
Half-Life	7.64 h	$\approx 8 - 12$ h	Yes	Within range.
LD50 / Acute Toxicity	45.1%	Very low toxicity	Yes	Correct qualitative match.
AMES Mutagenicity	Negative	Negative	Yes	Correct.
Carcinogenicity	No	No	Yes	Correct.
DILI Risk	No	No	Yes	Correct.
hERG Inhibition	Blocks hERG	No	No	False positive; could raise safety concerns.
Skin Reaction Risk	Causes reaction	Small but known risk	Yes	Correct directionally.
Lipophilicity (LogP)	2.54	2.9	Partial	Slight underestimation, but close.
Solubility	-3.53 log mol/L	-3.43 log mol/L	Yes	Accurate.

A.2 Rasagiline — Ensemble vs. Ground Truth

Table 3: Comparison between ensemble prediction and ground truth of Rasagiline.

Property	Ensemble Prediction	Ground Truth	Match?	Notes
Oral Bioavailability	76.34%	$\approx 36\%$	No	Significant overestimation.
HIA	97.35%	Complete/rapid absorption	Yes	Consistent classification.
P-gp Substrate	56.67%	Not a substrate	No	Incorrect classification.
Caco2 Permeability	$-4.85\,\mathrm{cm\,s^{-1}}$	Low permeability	Yes	Matches directionally.
BBB Penetration	83.13%	Yes	Yes	Correctly predicted.
PPBR	80.55%	$88 - 94\%$	Partial	Slight underestimation.
CYP1A2 Substrate	71.31%	Primary pathway	Yes	Accurate.
Half-Life	7.65 h	$0.6 - 2$ h	No	Model severely overestimates exposure duration.
AMES Mutagenicity	Not mutagenic	Negative	Yes	Correct.
DILI Risk	Cannot cause DILI	Low/unclear risk	Yes	Acceptable classification.
hERG Inhibition	Does not block	No inhibition	Yes	Correct.
Skin Reaction Risk	Does not cause	Rare hypersensitivity	Yes	Acceptable approximation.
Lipophilicity (LogP)	2.73	1.84	No	Overestimated lipophilicity.
Solubility	-4.35 log mol/L	-3.8 log mol/L	Partial	Ensemble predicts slightly lower solubility than reality.

A.3 Ambroxol — GCase docking prediction

Table 4: Comparison between ensemble prediction and ground truth for GCase–ambroxol binding prediction.

Prediction	Ground Truth	Prediction match?
• **Protein / structure:** GCase (GBA1), docked on **2NSX** (active-site reference). • **Pose:** Ligand sits **at the active-site mouth**, near catalytic **E235** (proximal to E340). • **Docking score:** $-6.9\ \mathrm{kcal\,mol^{-1}}$ (suggests low- to mid-μM affinity). • **Implication:** Consistent with an **active-site–proximal modulator**.	• **Mechanism: pH-dependent, mixed-type inhibition** (potent at neutral pH; weak to none at lysosomal pH). • **Potency:** About $K_i \approx 5\ \mu\mathrm{M}$ at $\mathrm{pH} \approx 7$, $\sim$ 20–30 $\mu\mathrm{M}$ at $\mathrm{pH} \approx 5.6$, and minimal/none at $\mathrm{pH} \leq 4.7$. • **Direct binding:** Global stabilization assays indicate **weak binding** at neutral pH ($K_d \sim 10^2\ \mu\mathrm{M}$) and **undetectable** at acidic pH (methodological and pH differences vs. enzyme kinetics). • **Structures: No ABX–GCase co-crystal**; **2NSX** is GCase with an active-site ligand (IFG) used as a docking template. Experimental footprint maps to **E235**-adjacent loops (e.g., around Tyr244/Ser237).	• **Site:** ✓ Yes — pose near **E235** matches the experimentally inferred binding region. • **Affinity magnitude:** ✓ For **neutral pH**, $-6.9\ \mathrm{kcal\,mol^{-1}} \approx \mu\mathrm{M}\ K_i$, aligning with kinetic data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims match methods/results focusing on an ensemble pipeline and a lead candidate; see Sections 1 (Table) and Results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Our results are entirely *in silico*; no synthesis or wet-lab validation has yet been performed. Docking scores depend on receptor preparation and scoring-function biases, and LLM-based MoA screening can inherit literature biases and hallucinations. ADMET predictions may not fully capture CNS-specific liabilities or idiosyncratic toxicity. Finally, manufacturability and IP novelty assessments require deeper expert and legal review.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No formal theorems are presented.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We detail stages, models, docking target (PDB: 3GXI), and criteria in Methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This submission only describes a framework

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Docking setup, targets, and filtering thresholds are described in Methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports point estimates (e.g., docking scores, accuracy) but does not include error bars or confidence intervals, nor does it specify variability sources or how uncertainties were computed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We will include compute specifics in the supplementary after de-anonymization if required.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Work involves in silico modeling only; no human/animal data.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: AI-enabled drug discovery for neurodegeneration could accelerate access to disease-modifying therapies, but it also raises concerns: (i) model misuse or overreliance on unvalidated predictions, (ii) unequal access to computational resources, and (iii) potential IP/ethics issues around training data. We encourage transparent reporting, reproducible pipelines, and responsible dissemination practices.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No model/dataset release in submission.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses existing assets (e.g., PDB structure 3GXI for GCase, AutoDock Vina for docking, and model/tooling such as TxGemma and FutureHouse/Stanford Biomni).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This submission does not release new assets (code, or data) at submission time, so the documentation requirement does not apply. If assets are released post-review, we will provide structured documentation via an anonymized package or URL.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM-based MoA screening and orchestration are described in Methods.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.