Explanations Aren't Just for Humans: XAI Boosts Agent-Driven Molecular Optimization

DS&AI - MLR

Andrea Hunklinger*[1,2], Fabian Krüger*[1,3,4], Adrian Wolny[1], Tim Adler[1], Igor Tetko[3], Santiago Villalba[1]

[1]Bayer AG, [2]University of Barcelona, [3]Helmholtz Munich, [4]Technical University Munich, *Equal contribution

Introduction

Generating molecules with optimized properties is central to drug discovery, where candidates must satisfy multiple **competing objectives** including potency, selectivity, and pharmacokinetics. The searchable **chemical space is vast**, estimated at ~ 10^{60} drug-like molecules, making exhaustive exploration impossible. Practical molecular design therefore proceeds iteratively: propose, evaluate, refine.[1] **Computational oracles** range from fast but less accurate methods to expensive but predictive approaches like protein–ligand co-folding.[2]

→ Reducing required evaluations enables use of higher-quality oracles

SEISMO treats optimization as an inference-time, sequential decision process within the LLM's context. Each oracle call immediately updates the state for the next proposal without batching. The agent conditions on:

- Full optimization trajectory (all prior molecules, scores, feedback)
- Natural-language task descriptions (leveraging LLM's chemical/biological priors)
- Structured explanatory feedback (sub-score decompositions, SHAP[3] attributions)

→ Increase feedback density:
single scalar → information-rich observations

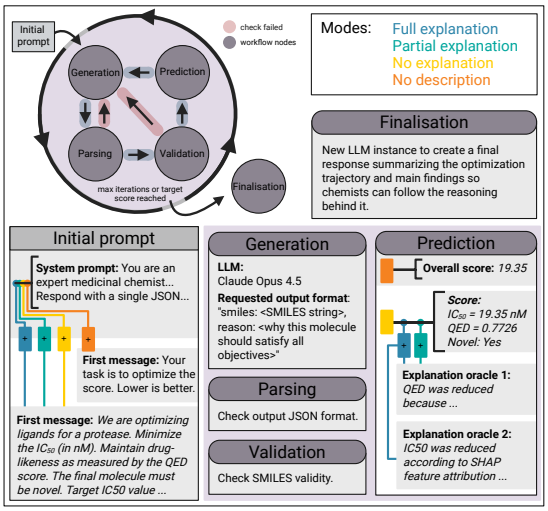


Figure 1: **SEISMO workflow overview**: The initial message and prediction feedback are modular and differ between tasks. *Italicized* text shows an example.

Benchmark performance

Model	∑ Top-1 AUC	∑ Best score	∑ Top-10 AUC
SEISMO	**19.85**	**21.18**	**20.24**
LLM prior	16.43	17.66	15.60
SEISMO (no description)	8.50	10.45	9.91
REINVENT[5]	7.06	8.06	10.68
LICO[6]	-	-	11.71

Table 1: Area under the optimization curve (AUC) using the top-k best scores found up to each oracle call. Values are summed over all 23 optimization tasks from PMO[4]. Colors show number of oracle calls : **blue** (50, values fixed afterwards), **orange** (10,000), and **yellow** (1,000).

- **Strong performance** of SEISMO: near-optimal scores (≥0.9) on 16/23 tasks within 50 oracle calls on PMO benchmark[4]
- Previously found solutions may overlap with training data, meaning benchmark performance could reflect **memorization** rather than true optimization
- Planned to confirm extrapolative ability and test designs **experimentally**

Co-folding oracle

- **Goal:** Optimize small molecules for high predicted binding to a fixed protein target using a more computationally expensive oracle: Boltz-2 [2]
- **Explanation:** All residues within 5 Å of the ligand (heavy-atom distance) + Boltz-2 confidence scores
- Additional context to localize binding site and interactions seemed more effective than sequence information alone
- **Limitation:** Boltz-2 binding probability not directly translatable to experimental binding affinities → FEP-simulations or lab experimentation next

Target	Difficulty	Rationale	Result
SARS-CoV-2 Mpro	easy	many known binders; likely covered in LLM training data	Near-reference predicted binding strength achieved within first iterations
TRIB2 pseudo-kinase	medium	limited known reference ligands	exceeds reference in predicted binding probability; Boltz-2 possibly unreliable for this target
MGYP001550541752	hard	no known binders; high-confidence 3D structure; reduced memorization effect	SEISMO outperforms conventional generative methods before their first batch update

Table 2: Molecular optimization targets and results using Boltz-2 binding probability as the oracle.

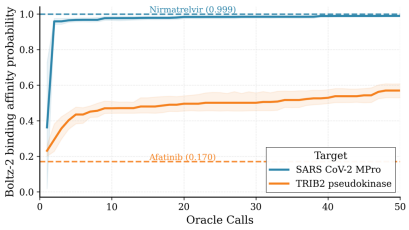


Figure 3: Best-so-far Boltz-2 predicted binding probability over 50 iterations vs reference ligands (Nirmatrelvir for Mpro (**easy**) and Afatinib for TRIB2 (**medium**)). SEISMO matches or exceeds references in predicted binding.

Can we make Boltz-2 more explainable? Domain-adaptive XAI could unlock the next leap.

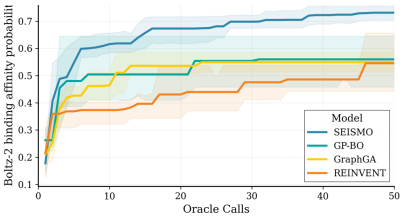


Figure 4: Best-so-far Boltz-2 binding affinity probability over oracle calls for the **hard** target. SEISMO outperforms REINVENT[5], and other batch- and population-based methods (GraphGA[7], and GP-BO[8]), which were added as traditional generative baselines.

Adding explanations

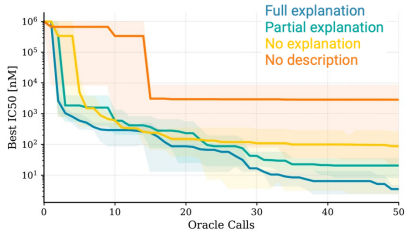


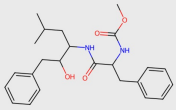
Figure 2: Best predicted IC_{50} (log scale, nM) versus oracle calls. Richer explanatory feedback progressively improves sample efficiency. Shaded regions indicate 95% confidence intervals.

- **Task**: Find novel, drug-like molecules (QED ≥ 0.6) with minimized predicted inhibitory constant (IC_{50}) for SARS-CoV-2 main protease (Mpro) inhibition
- **Oracle**: XGBoost regressor trained on MACCS fingerprints. Explanations extracted using SHAP[3]
- **Key finding**: Adding explanations to the scalar score further enhances sample efficiency. Variants with richer feedback increasingly outperform those with less information as more oracle calls are made
- **Results**: Within 50 iterations: 2.5 nM predicted IC_{50} and below 1 nM in 150 iterations

Example

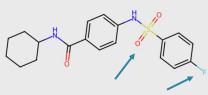
Iteration 1: *"Starting with a peptidomimetic scaffold commonly found in protease Inhibitors"*

Predicted IC_{50}: **94377.57** nM
QED: **0.5596**



> **115-fold** improvement in one iteration

Iteration 2: *"The previous molecule had too many rotatable bonds, high MW, and multiple HBD hurting QED. SHAP analysis shows adding sulfur, halogens, and C=O groups could improve IC50. I'm designing a more rigid, smaller molecule with a sulfonamide group, fluorine substituent, and fewer rotatable bonds while maintaining key pharmacophoric features for protease inhibition."*



Predicted IC_{50}: **810.56** nM
QED: **0.8357**

IC50 prediction using SHAP: (iteration 1)
Top 10: highest attribution values

Attribution	MACCS fingerprint value	MACCS key	SMARTS substructure
0.234509	0	66	[#6]~[#6](~[#6])(~[#6])~*
0.074470	1	55	[#8]~[#16]~[#8]
0.072645	0	81	[#16]~*(~*)~* ← Sulfur
0.060045	0	154	[#6]=[#8] ← C=O
0.058869	1	150	*!@*@*!@*
0.049210	0	101	<8M ring>
0.046236	0	87	[F,Cl,Br,I]!@*@* ← halogens
0.036998	0	84	[NH2]
0.032856	1	137	[!C;!c;R]
0.030273	1	107	[F,Cl,Br,I]~*(~*)~*

QED: (iteration 1)
Weighted log contributions (closer to 0 = better, negative values hurt score):

MW: -0.459 (value=412.53, optimal=281-332 Da)
ALOGP: -0.003 (value=3.09, optimal=1.95-3.57)
HBA: -0.006 (value=4, optimal=2-3)
HBD: -0.592 (value=3, optimal=1)
PSA: -0.013 (value=87.66, optimal=40-69 \u0172)
ROTB: -1.009 (value=10, optimal=2-4)
AROM: -0.031 (value=2, optimal=1-2)
ALERTS: -0.163 (value=0, optimal=0)

References:

[1] Ghiandoni, G. M., et al. "Augmenting DMTA using predictive AI modelling at AstraZeneca." *Drug discovery today* 29.4 (2024): 103945.
[2] Passaro, S., et al. "Boltz-2: Towards accurate and efficient binding affinity prediction." *BioRxiv* (2025).
[3] Lundberg, S. M., and Lee, S. I., "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
[4] Gao, W., et al. "Sample efficiency matters: a benchmark for practical molecular optimization." *Advances in neural information processing systems* 35 (2022): 21342-21357.
[5] Olivecrona, M., et al. "Molecular de-novo design through deep reinforcement learning." *Journal of cheminformatics* 9.1 (2017): 48.
[6] Nguyen, T., and Grover, T., "Lico: Large language models for in-context molecular optimization." *arXiv preprint arXiv:2406.18851* (2024).
[7] Jensen, J. H. "A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space." *Chemical science* 10.12 (2019): 3567-3572.
[8] Tripp, A., et al. "A fresh look at de novo molecular design benchmarks." *NeurIPS 2021 AI for Science Workshop*. (2021).

Learn more about it:

GitHub
FabianKruger/molecule-optimization-agent

arXiv
https://arxiv.org/abs/2602.00663

Andi.hunklinger@bayer.com
@ahunklinger.bsky.social