Revealing Token-Level Importance in Conditional Molecular Design Through Kullback–Leibler Divergence

Arthur Cerveira, Ulisses B. Corrêa

Programa de Pós-graduação em Computação (PPGC)
Hub de Inovacão em Inteligência Artificial (H2IA)
Centro de Desenvolvimento Tecnológico (CDTec),
Universidade Federal de Pelotas (UFPel), Brazil

{aacerveira,ulisses}@inf.ufpel.edu.br

Abstract. *Conditional molecular design using transformer-based models can accelerate drug discovery, but their black-box nature limits interpretability. We propose a method to explain these models by quantifying the influence of property conditions on the generative process. Our approach uses the Kullback–Leibler Divergence to measure the difference between conditional and unconditional output distributions at each generation step. This allows us to identify the token-level importance of a sequence for specific desired conditions.*

1. Introduction

Conditional molecular design is a key approach in computer-aided drug discovery (CADD) for generating new molecules with specific properties [Mak et al. 2023]. By conditioning the generation process on characteristics like binding affinity or desired physicochemical properties (e.g. solubility, stability, lipophilicity, etc.), these models can accelerate the identification of viable drug candidates [Wang et al. 2023]. Many deep learning architectures have been applied to this task, with transformer-based models being effective due to their ability to handle long-range dependencies in molecular sequences [Zhang et al. 2024].

In parallel, explainable artificial intelligence (XAI) is being applied to drug discovery as a means to interpret the predictions of complex models and provide chemists with useful insights [Alizadehsani et al. 2024]. While prior work has explored conditional generation architectures and applied explainability techniques to CADD, the integration of explainability into conditional molecular design remains unexplored. In particular, no methods currently leverage conditional molecular design architectures to produce explainable outputs that reveal the relative importance of each generated token with respect to the conditioning objective. Such token-level importance can highlight chemically meaningful substructures and molecular scaffolds, thereby providing a deeper understanding of the design process.

To address this, we propose a method to interpret the generative process of a conditional transformer model. Our approach uses the Kullback–Leibler (KL) Divergence to compute the difference between the unconditional and conditional probability distributions [Kullback and Leibler 1951]. This divergence serves as a proxy for how much the target property influences the generation of each token. By applying this at each step, we can quantify the impact of the condition on the resulting molecular structure, providing a necessary layer of transparency into the model's decision-making process.

2. Methods

While the proposed method is architecture-agnostic, we apply it to CoMPO-GPT, a transformer-based decoder conditioned on an aggregated latent representation of desired properties [Cerveira et al. 2025]. The model's ability to optimize for multiple properties simultaneously makes it an practical test case for exploring explainability in multi-objective drug discovery. For input, molecules are represented as simplified molecular-input line-entry system (SMILES) strings and the conditions are represented as a set of tokens $\{p_1, p_2, ..., p_n\}$ in a vocabulary.

The conditioning process begins with an embedding layer, where each desired property p_i (e.g., high binding affinity, low toxicity) is mapped to a learned representation e_{p_i}. For multi-objective optimization, a pooling layer aggregates the embeddings for multiple properties into a single, unified representation, e'_{multi}:

$$e'_{\text{multi}} = \text{pool}(e_{p_1}, e_{p_2}, \ldots, e_{p_n}) \tag{1}$$

where pool$(\cdot)$ can be a function such as mean, sum, or max pooling. This aggregated representation is then passed through a projection layer to match the dimensionality of the decoder's hidden states, resulting in the final multi-objective representation, e_{multi}:

$$e_{\text{multi}} = \text{proj}(e'_{\text{multi}}) \tag{2}$$

This representation conditions the decoder via a cross-attention mechanism. The decoder's hidden states serve as the query (Q), while the multi-objective representation e_{multi} provides the keys (K) and values (V):

$$Q = H_{\text{dec}} \cdot W_q \tag{3}$$
$$K = e_{\text{multi}} \cdot W_k \tag{4}$$
$$V = e_{\text{multi}} \cdot W_v \tag{5}$$

where H_{dec} denotes the decoder hidden states, and W_q, W_k, and W_v are learnable weight matrices. Additionally, the e_{multi} embedding is summed with the input embeddings of the molecule tokens, ensuring the conditioning signal propagates through all layers of the decoder.

To determine the importance of the set of conditions, we measure the change in the model's output distribution using the KL Divergence. We compare the conditional probability distribution over the next token y_t, $P_c(y_t) = P(y_t|y_{<t}, e_{\text{multi}})$, with the unconditional distribution, $P_u(y_t) = P(y_t|y_{<t}, \emptyset)$, where a default unconditional token $\emptyset$ is applied instead. The KL divergence is computed as:

$$D_{KL}(P_c||P_u) = \sum_{y_t \in V} P_c(y_t) \log \frac{P_c(y_t)}{P_u(y_t)} \tag{6}$$

where V is the vocabulary of all possible tokens. A large KL divergence value indicates that the specified properties in e_{multi} significantly alters the probability of the next token, suggesting that the next generated token is highly relevant for guiding the generation process toward the desired properties $\{p_1, p_2, ..., p_n\}$.

This process is repeated for each generated token until an end-of-sequence ($[EOS]$) token is produced, providing a step-by-step analysis of the conditional influence, as illustrated in Figure 1. A key advantage of this method is its ability to directly leverage CoMPO-GPT multi-objective mechanism, allowing for the simultaneous identification of token importance across multiple properties, a task that is not as straightforward in other explainability approaches that often focus on single-condition explanations.

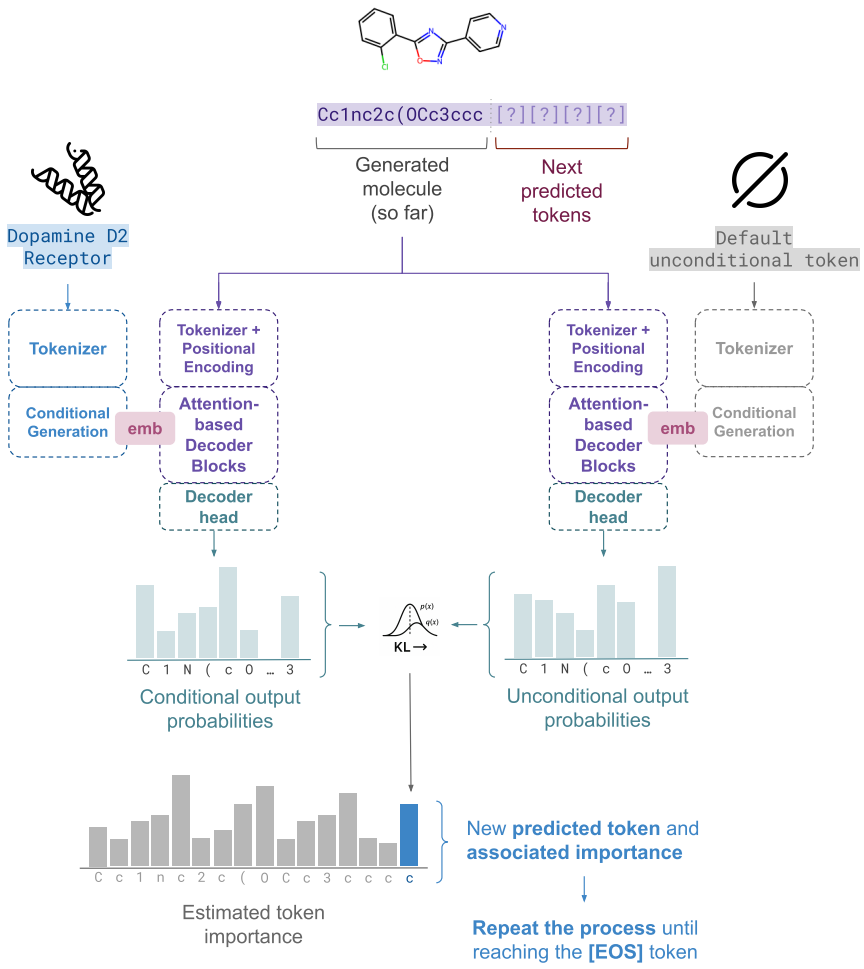

Figure. 1. An overview of the proposed explainability method. Token importance is calculated via the KL Divergence between conditional and unconditional generation steps. The example shows the identification of key tokens when designing a molecule active for the Dopamine D2 Receptor.

3. Experiments and Expected Results

The described process for obtaining token importance for a single or multiple properties has been implemented. An illustrative example of this analysis is presented in Figure 2, which shows the token-level importance scores during a conditional generation task. For future work, we intend to conduct experiments to further validate and explore this method. First, we will analyze how perturbations on the most important tokens (identified by high KL divergence) affect the predicted property values of the generated molecules. This will

help confirm that the identified tokens are indeed critical for achieving the desired conditions. Second, we plan to compare the results of our explainability method with other approaches in the literature to benchmark its performance and highlight its advantages.

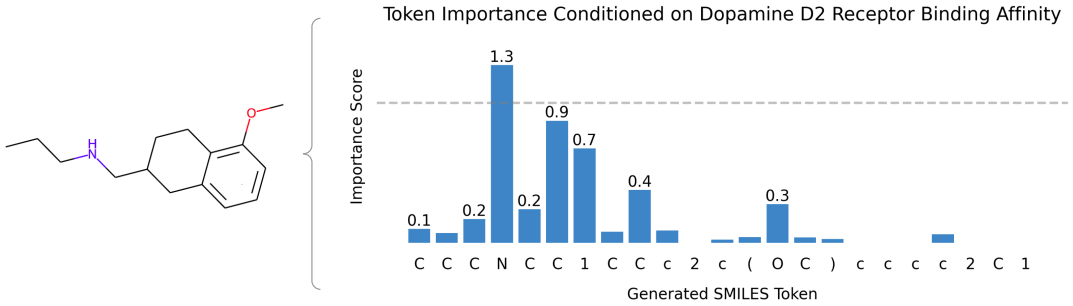

Figure. 2. Token importance for a molecule optimized for binding affinity.

4. Conclusions

We introduced a novel method for interpreting conditional molecular design models using KL divergence to quantify the influence of property conditions. This approach provides a direct way to understand how the model prioritizes different objectives during generation. Future work will focus on validating these explanations through perturbation studies and benchmarking against other methods. We also intend to assess how this method generalizes to other textual- and graph-based molecular representations.

References

Alizadehsani, R., Oyelere, S. S., Hussain, S., Jagatheesaperumal, S. K., Calixto, R. R., Rahouti, M., Roshanzamir, M., and De Albuquerque, V. H. C. (2024). Explainable artificial intelligence for drug discovery and development: A comprehensive survey. *IEEE Access*, 12:35796–35812.

Cerveira, A., Kremer, F., Gomes, G., and Correa, U. (2025). Compo-gpt: Cross-attention conditioning for multi-target molecular design in generative models. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Mak, K., Wong, Y., and Pichika, M. (2023). Artificial intelligence in drug discovery and development. In Hock, F. and Pugsley, M., editors, *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*. Springer, Cham.

Wang, Y., Zhao, H., Sciabola, S., and Wang, W. (2023). cmolgpt: A conditional generative pre-trained transformer for target-specific de novo molecular generation. *Molecules*, 28(11):4430.

Zhang, Y., Liu, C., Liu, M., Liu, T., Lin, H., Huang, C.-B., and Ning, L. (2024). Attention is all you need: utilizing attention in ai-enabled drug discovery. *Briefings in Bioinformatics*, 25(1):bbad467.