Nahal, Yasmine; Heinonen, Markus; Kabeshov, Mikhail; Janet, Jon Paul; Nittinger, Eva; Engkvist, Ola; Kaski, Samuel

Towards Interpretable Models of Chemist Preferences for Human-in-the-Loop Assisted Drug Discovery

Published in:
AI in Drug Discovery - 1st International Workshop, AIDD 2024, Held in Conjunction with ICANN 2024, Proceedings

DOI:
10.1007/978-3-031-72381-0_6

Published: 01/01/2025

Document Version
Publisher's PDF, also known as Version of record

Towards Interpretable Models of Chemist Preferences for Human-in-the-Loop Assisted Drug Discovery

Yasmine Nahal[1(✉)], Markus Heinonen[1], Mikhail Kabeshov[2], Jon Paul Janet[2], Eva Nittinger[3], Ola Engkvist[2,4], and Samuel Kaski[1,5]

[1] Department of Computer Science, Aalto University, Espoo, Finland
{yasmine.nahal,markus.heinonen,samuel.kaski}@aalto.fi
[2] Molecular AI, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
{mikhail.kabeshov,jon.janet,ola.engkvist}@astrazeneca.com
[3] Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
eva.nittinger@astrazeneca.com
[4] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
[5] Department of Computer Science, University of Manchester, Manchester, UK

Abstract. In recent years, there has been growing interest in leveraging human preferences for drug discovery to build models that capture chemists' intuition for *de novo* molecular design, lead optimization, and prioritization for experimental validation. However, existing models derived from human preferences in chemistry are often black-boxes, lacking interpretability regarding how humans form their preferences. Enhancing transparency in human-in-the-loop learning is crucial to ensure that such approaches in drug discovery are not unduly affected by subjective bias, noise or inconsistency. Moreover, interpretability can promote the development and use of multi-user models in drug design projects, integrating multiple expert perspectives and insights into multi-objective optimization frameworks for *de novo* molecular design. This also allows for assigning more or less weight to experts based on their knowledge of specific properties. In this paper, we present a methodology for decomposing human preferences based on binary responses (like/dislike) to molecules essentially proposed by generative chemistry models, and inferring interpretable preference models that represent human reasoning. Our approach aims to bridge the gap between human-in-the-loop learning and user model interpretability in drug discovery applications, providing a transparent framework that elucidates how human judgments can shape molecular design outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72381-0_6.

D.-A. Clevert et al. (Eds.): AIDD 2024, LNCS 14894, pp. 58–70, 2025.
https://doi.org/10.1007/978-3-031-72381-0_6

Keywords: Human-in-the-loop machine learning · Feature decomposition · User modelling · Interpretability · De novo molecular design

1 Introduction

Designing effective molecule scoring functions for drug discovery is a highly complex and multifaceted challenge. This complexity arises from the need to balance multiple objectives, such as potency, selectivity, toxicity and synthetic accessibility (SA), each of which must be optimized simultaneously. Traditional computational methods often struggle to integrate these diverse factors into a single, coherent scoring function, making the discovery process both time-consuming and uncertain. The dynamic nature of biological systems and the vast chemical space further complicate the task, requiring innovative approaches to accurately evaluate molecular efficacy beyond conventional manually-engineered scoring functions.

Human-in-the-loop assisted drug discovery offers a promising solution by incorporating the expertise and intuition of chemists directly into the computational workflow. Unlike static, manually-defined scoring functions, human-in-the-loop learning approaches allow for real-time adjustments based on expert knowledge and evolving insights. This dynamic interaction enables a more nuanced evaluation of candidate molecules, leveraging human judgment to guide the discovery process more effectively. By integrating human expertise, data-driven methods can adapt to new information and improve the relevance and accuracy of molecular predictions.

Several models have been developed to harness human input in drug discovery. Notable among these are the works of Sundin et al. [15] and MolSkill [5]. While Sundin et al. focus on dynamically learning a scoring function from binary human preference responses on proposed designs, MolSkill authors train a neural network using pairwise comparisons between molecules to infer user preferences. These models represent significant advancements in incorporating human preferences into drug discovery, yet they often operate as black-box systems with limited transparency.

The lack of interpretability in these user models is a critical concern. Black-box user models can obscure the rationale behind chemist intuition, making it difficult to trust and validate the outcomes. This may represent a bottleneck to the effective integration of human expertise into the drug discovery process.

A previous study by Kutchukian et al. [10] has shown that medicinal chemists simplify the complex task of identifying promising compounds by focusing on a subset of parameters, despite the complexity involved. Moreover, the study highlighted discrepancies between chemists' reported decision criteria and the actual parameters that influence their choices, emphasizing the need for more transparent and interpretable user models in drug discovery.

To address these challenges, we propose inferring interpretable user models by decomposing observed preference data into meaningful features or molecular

descriptors. Feature decomposition [12] is a supervised learning strategy with the potential to efficiently dissect user preference data and understand the underlying factors influencing their decisions. This approach aligns with related work in feature decomposition, which has been successfully applied in various fields such as social science to enhance transparency and robustness of human behavioral models [8]. By adopting a similar strategy, we aim to create interpretable user models of chemist intuition that can be used for molecular design, optimization or experiment prioritization, which can later be integrated in drug discovery pipelines without the need for direct human intervention.

In this paper, we propose a methodology for decomposing human preferences, presented as binary responses (i.e., like/dislike) to molecules either proposed by generative chemistry models or from existing chemical libraries. Our approach seeks to bridge the gap between user modelling and model interpretability for human-in-the-loop assisted drug discovery. By providing a transparent framework, we aim to elucidate how human judgments shape molecular design outcomes, ultimately encouraging the reliance on user models and hybrid machine-user models as scoring functions.

2 Methodology

We consider a setting where a user provides a binary response $y \in \{0, 1\}$ to a molecular design $\mathbf{x}$, reflecting their preference for the design. We assume a response model

$$y \sim \text{Ber}\left(\text{sigmoid}\left(\mathbf{w}^T g(\mathbf{x})\right)\right) \tag{1}$$

where the user's binary response comes from a Bernoulli distribution, with the probability given by a sigmoidal function of $\mathbf{w}^T g(\mathbf{x})$. The function $g(\mathbf{x}) \in \mathbb{R}^D$ represents the features considered by the user in their decision, and $\mathbf{w} \in \mathbb{R}^D$ represents a linear weighting of these features. For example, the user might consider the following descriptors

$$g(\mathbf{x}) = (\text{synthetizability}(\mathbf{x}), \text{solubility}(\mathbf{x}), \text{patentability}(\mathbf{x}), \text{activity}(\mathbf{x}))$$

and weigh them by $(0.5, 0.3, 0.4, 0.7)$ respectively. Given that users may make errors in their mental evaluation of descriptors, we assume g is an approximation of some underlying true function $g^\star$

$$g(\mathbf{x}) = g^\star(\mathbf{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}^2)) \tag{2}$$

For instance, a user might misjudge the solubility of a molecule.

Furthermore, we consider that each user has a different labelling function. Let j index the labeller, and define its labelling function as

$$y_j \sim \text{Ber}\left(\text{sigmoid}\left(\mathbf{w}_j^T g_j(\mathbf{x})\right)\right) \tag{3}$$

Thus, by querying a collection of users $(1, \ldots, J)$ about a single molecule, we obtain a set of binary responses $(y_1, \ldots, y_J)$, each stemming from different preference functions $\mathbf{w}_j^T g_j(\mathbf{x})$.

We assume a dataset of binary 'votes' $\mathbf{Y} = (y_{ij}) \in \mathbb{R}^{N \times J}$, where y_{ij} is the label of the i-th molecule from the j-th user.

Our goal is to infer the mental variables $\{\mathbf{w}_j, g_j\}$, but this task is unrealistic in its current form, since we have no direct knowledge of the mental descriptors g_j or the weightings $\mathbf{w}_j$ of the users.

Instead, we simplify the problem by assuming that all experts share the same descriptors. More precisely, we assume that the set of descriptors used by any expert is the union of all expert descriptors, with unused descriptors showing as zeros in $\mathbf{w}$. We thus reformulate the model as

$$y_j \sim \text{Ber}\left(\text{sigmoid}\left(\mathbf{w}_j^T g(\mathbf{x})\right)\right) \tag{4}$$

Given that fitting binary outcome variables directly might not yield a closed-form solution, we consider using a probit link function instead of the logit (sigmoid) function. The probit link can simplify the sampling process, resulting in a straightforward Gibbs sampler when only the weights are learned. This modification is expressed as:

$$y_j \sim \text{Ber}\left(\Phi\left(\mathbf{w}_j^T g(\mathbf{x})\right)\right) \tag{5}$$

where Φ denotes the cumulative distribution function (CDF) of the standard normal distribution.

Finally, our problem is to infer the posterior

$$p(\mathbf{W}, g \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \mathbf{W}, g) p(g) \prod_j p(\mathbf{w}_j) \tag{6}$$

where all weights $\mathbf{w}_j$ share the same prior. We aim to determine the mental descriptors g and the user preferences $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_J) \in \mathbb{R}^{J \times D}$. Prior knowledge can be incorporated into this problem. First, the preferences $\mathbf{w}_j$ are assumed to be sparse, for instance, by employing a Horseshoe prior $p(\mathbf{w}_j) = \mathcal{HS}(\mathbf{w}_j)$. Second, the descriptor function g can be fixed to a dictionary of known descriptors (e.g., from chemoinformatics software) or can be the feature embedding of a molecular deep learning network (e.g., MolBert). A more advanced approach would be to treat these as the prior mean functions and infer slight fine-tuning of them.

Given the expense of human evaluations, even if this model does not have a closed-form solution, MCMC methods such as Hamiltonian Monte Carlo (HMC) are likely feasible options. These methods can handle the complex posterior distributions involved in our problem and provide robust estimates of the parameters.

In summary, while a closed-form solution would be ideal for computational simplicity, the use of probabilistic methods like HMC offers a practical and effective route for inference in our setting. We will implement and compare both logit

and probit models, evaluating their performance using real-world data to ensure the robustness and applicability of our approach.

3 Experiments

We conducted experiments to infer interpretable user models for chemist preferences in molecular design using Bayesian inference with a Stan model. Our approach involved querying experts to provide binary preference responses (*like*/*dislike*) for a dataset of molecules represented by molecular descriptors. The model assumes that each expert's preference is influenced by a weighted combination of these descriptors.

3.1 Experimental Setup

Data Collection. We collected a dataset consisting of binary responses ($\mathbf{Y}$) from $J = 3$ experts for $N = 150$ molecules generated using the molecular design tool REINVENT [3]. The experts were asked to rate those molecules based on how much they align with the molecular design objective of producing novel binders for the Dopamine receptor D2 (DRD2). Feedback from experts was collected in real-time through the Metis interface [9] by sampling (after a defined number of reinforcement learning steps) molecules from the generative chemistry model implemented in REINVENT. This generative chemistry model was trained to maximize predicted DRD2 probabilities by a Quantitative Structure-Activity Relationship (QSAR) model. A screenshot of the interface is provided in Figure S6, where structures of generated molecules are displayed alongside their DRD2 activity probabilities. The experts were asked to express, on a scale from 0 to 100, how much they liked the proposed DRD2 binders. Expert scores were transformed into binary labels using a threshold value of 50 and included in the initial training dataset of the DRD2 QSAR model, which was then used to guide the generation of subsequent DRD2 binders by REINVENT. This iterative process ensured that the generative chemistry model continually improved in alignment with expert preferences. All three participating experts are co-authors of this manuscript.

For the set of evaluated molecules, we calculated molecular descriptors using RDKit [1], which include the molecular weight (`MolWt`), number of rotatable bonds (`NumRotaBonds`), the logarithm of the octanol-water partition coefficient or LogP (`MolLogP`), the number of aromatic rings (`NumAromRings`), the number of hydrogen bond acceptors (`HBA`) and donors (`HBD`), the topological surface area (`TPSA`), and the structural alerts or undesirable substructures according to the Quantitative Estimate of Drug-likeness (QEDAlerts) [2]. For the latter, we modified the standard QED implementation in RDKit by setting the weights for all other properties (`MolWt`, `MolLogP`, `HBA`, `HBD`, `TPSA`, `NumRotaBonds`, `NumAromRings`) to 0 and only keeping the weight for the presence of undesirable substructures to 1. This ensures that the QED score solely reflects the presence of structural alerts. Additionally, we used the SA score developed by Ertl et

al. [7], as well as the probability of DRD2 bioactivity according to the classifier developed by Olivecrona et al. [13], as descriptors that can explain the user preference responses.

We analyzed the Pearson correlations among the molecular descriptors used for this study (Figure S7). Notably, `MolLogP` shows the strongest positive correlation (0.78) with `TPSA`. All correlations, ranging from -0.63 to 0.78, are indicative of meaningful relationships between descriptors that can enhance model accuracy and interpretability.

Model Specification. The Bayesian model was implemented using the Stan probabilistic programming language [4]. The model included:

- Parameters:
  - τ: Global scale parameter controlling the overall sparsity of weights assigned to the molecular descriptors.
  - λ_j: Local scale parameters for each expert j.
  - $\mathbf{w}$: Preference weights matrix, where each column represented the weights for one expert across all descriptors.
- Priors:
  - $\tau \sim \text{Cauchy}(0, \tau_0)$: Cauchy prior for global shrinkage.
  - $\lambda_j \sim \text{Cauchy}(0, 1)$: Cauchy priors for local shrinkage.
  - $\mathbf{w} \sim \text{Normal}(0, \lambda_j \cdot \tau)$: Normal priors for weights adjusted by local scales.
- Likelihood:
  - $\mathbf{Y}_{nj} \sim \text{Bernoulli}(\text{logit}(\mathbf{X} \cdot \mathbf{w}_{\cdot j}))$: Likelihood of expert j's response based on the linear combination of molecular descriptors weighted by $\mathbf{w}_{\cdot j}$.

3.2 Implementation

The Stan model was compiled and fitted to the data using Hamiltonian Monte Carlo (HMC) sampling (2000 iterations, 2 Markov chains with a maximum tree depth of 15 and the parameter `adapt_delta` set to 0.99). This approach enabled us to approximate the posterior distribution of parameters $\mathbf{w}$ and λ_j, which represent the preference weights and local scale parameters, respectively.

Since the dataset is already very small (due to limited resource availability for human data collection), we die not split it into training and testing, and chose to fit the model to the entire dataset instead to reach the highest accuracy.

Convergence diagnostics, including the $\hat{R}$ statistic and the trace plots, were performed to assess the model's convergence and ensure reliable inference across multiple chains. The $\hat{R}$ statistic, also known as the potential scale reduction factor, is a convergence diagnostic used to assess whether the Markov chains in the MCMC sampling have converged to the target distribution (i.e., response labels). Specifically, it compares the variance within chains to the variance between chains. A value close to 1 (typically $\hat{R} < 1.1$) indicates convergence, which is what we have observed with our model. Trace plots are visual representations that show how the Markov chain samples evolve over iterations, allowing to diagnose issues like non-stationarity and mixing problems. Our trace plots (Figure S5) show that the Markov chains have mixed properly (low divergence).

3.3 Benchmark

We compared our model against a non-probabilistic logistic regression (LogReg) and a Random Forest Classifier (RFC), implemented using the Scikit-learn package [14]. The purpose is to demonstrate that our model is more transparent than its non-probabilistic counterparts, allowing for direct interpretability of the reasoning process behind the human preferences, in addition to a reasonable classification accuracy. The same set of molecular descriptors described in Sect. 3.1 was used to train the LogReg and RFC models on the 150 human-rated molecules by each expert, individually. The classification accuracy scores (i.e., percentages of correctly classified molecules into liked or disliked) were calculated for each individual user model, then the average accuracy scores were reported. To assess the interpretability of the RFC models, Shapely values for tree-based algorithms were computed [11]. For LogReg models, we analyzed feature importance.

4 Results

4.1 Interpretability of Human Preferences

We consider that a model is able to accurately interpret human preferences based on how the participating experts described their reasoning. Our Stan model effectively deciphered the human reasoning behind the preference dataset for the DRD2 binders. The learned descriptor contributions are weights are illustrated in Fig. 1.

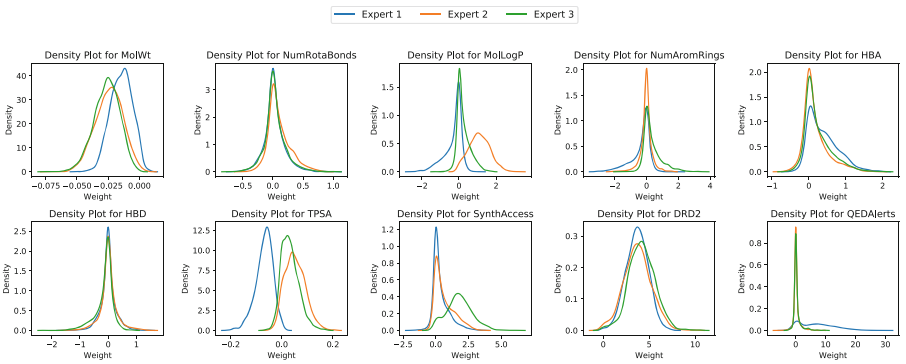

Fig. 1. Density plots of the molecular descriptor weights learned by the Stan model for each expert. Each subplot corresponds to a different descriptor, with density curves representing the weight distributions for Expert 1 (blue), Expert 2 (orange) and Expert 3 (green). These plots illustrate the variations in how each expert weighted the descriptors, reflecting their individual preferences and reasoning. Notably, the plots show that the DRD2 bioactivity descriptor was consistently important across all experts, while other descriptors such as `MolLogP`, `SynthAccess` and `QEDAlerts` had varying levels of importance depending on the expert. (Color figure online)

When asked to describe their personal experiences from interacting with DRD2 binders generated using REINVENT, Expert 1 highlighted their focus on the structural characteristics of the generated molecules, rejecting many at the initial stages of the interaction because they appeared too different and "odd" compared to known DRD2 actives. This was accurately captured by the Stan model, showing greater correlation between Expert 1 preferential feedback and the `QEDAlerts` descriptor (Fig. 2). The focus of Expert 1 on the presence of structural alerts in the generated DRD2 binders has led to the generation of more drug-like molecules, which can be observed through an increased QED score (Table S1 where molecular generation performance metrics are reported for top-scoring DRD2 binders generated by REINVENT after incorporating expert preference feedback).

Conversely, Expert 3 described that they were more concerned with the SA of the generated DRD2 binders. This preference was well captured by the higher estimated weights for the `SynthAccess` descriptor (Fig. 1) and correlation between Expert 2 preferential feedback and `SynthAccess` (Fig. 2). Expert 2 described that they rated the molecules based on how much they liked them as a lead, aiming to select molecules that would be synthesizable, stable and with reasonable lipophilicity to maximize their chance for being made and tested. The Stan model's learned weights revealed that Expert 2 indeed prioritized the molecular LogP followed by synthetic accessibility, as indicated by the higher weights for those descriptors (Fig. 1) and stronger correlation with `MolLogP` (Fig. 2). Moreover, a higher percentage of lead-like compounds according to the rule of three (RO3) [6] for molecular LogP was identified based on feedback from Expert 2 (Table S1). Expert 2 showed similarities with Expert 3 in their reasoning regarding DRD2 binders: they both acknowledged not having any particular knowledge of the target or known binders.

Interestingly, the weights for the DRD2 bioactivity descriptor were high for all three experts, indicating that the model successfully captured that the preference feedback was related to the rating of DRD2 binders. These findings are consistent with the descriptions provided by the experts upon the completion of the interaction exercise, validating the model's ability to interpret and reflect their reasoning accurately.

The interpretability analysis from the RFC models also highlighted the importance of the DRD2 activity descriptor in explaining user preference feedback 3. For Expert 1, the RFC models accurately captured their preference for more complex molecular structures but did not fully reflect their reliance on the presence of structural alerts that could undermine drug likeness. For Experts 2 and 3, `MolLogP` and `SynthAccess` were correctly identified as important descriptors in explaining their feedback. However, `MolWt` was also identified as significantly important, though it was not explicitly emphasized in the expert feedback. Therefore, we consider the RFC models' interpretations to be close to the Stan model's performance, with the latter being the most aligned with the expert descriptions Fig. 3.

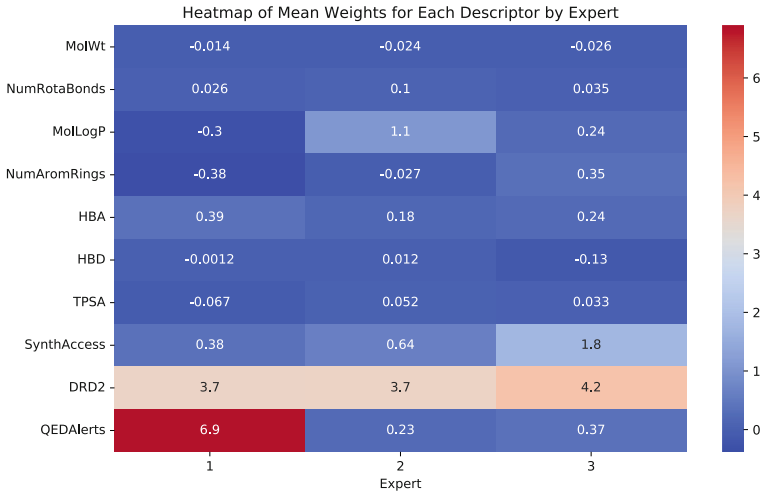

Fig. 2. Correlation between each expert and molecular descriptors according to the fitted Stan model. Heatmap matrix showing the relationship between each expert's feedback and the molecular descriptors used for fitting the Stan model. The matrix highlights a higher correlation between Expert 1 and `QEDAlerts`, Expert 2 and `MolLogP`, and Expert 3 and `SynthAccess`. All experts show a high correlation with the DRD2 activity descriptor, indicating its importance to explain the expert preferential responses.

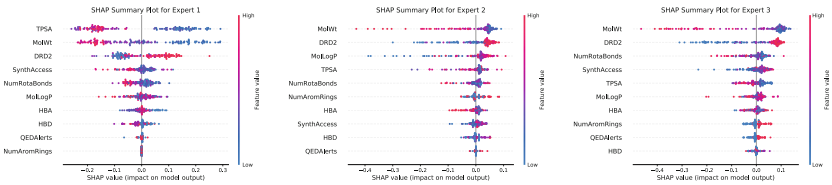

Fig. 3. SHAP summary plots for the Random Forest Classifier predictions, illustrating the importance of various molecular descriptors for each expert. The plots provide a visual explanation of how each descriptor contributes to the model's predictions, with distinct patterns emerging for each expert that align with their feedback preferences.

Conversely, the LogReg models provided the least accurate interpretation of descriptor importance. They failed to clearly capture Expert 1's emphasis on structural alerts and Expert 3's focus on SA for the generated DRD2 binders Fig. 4. This suggests that while the LogReg models can provide some insights, they are not as reliable as the Stan and RFC models in reflecting the experts' reasoning.

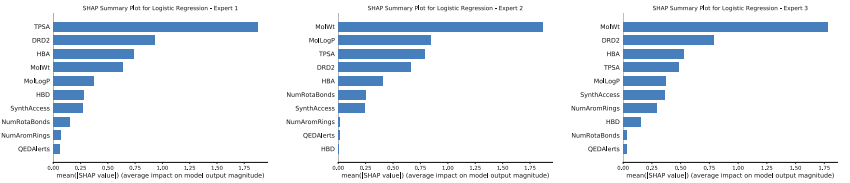

Fig. 4. SHAP summary plots for the Logistic Regression predictions, depicting the significance of molecular descriptors in determining the experts' feedback. These plots highlight the differences in descriptor importance across experts and provide insight into the Logistic Regression model's interpretation of the data.

4.2 Accuracy in Predicting Human Preferences

We compared our model against a non-probabilistic logistic regression (LogReg) and a Random Forest Classifier (RFC), implemented using the Scikit-learn package [14]. The primary goal of this comparison is to demonstrate the interpretability of our model in contrast to its non-probabilistic counterparts, while also showcasing its reasonable classification accuracy. The same molecular descriptors were used to train the LogReg and RFC models on the 150 human-rated molecules by each expert, individually. The classification accuracy scores (i.e., percentages of correctly classified molecules into liked or disliked) were calculated for each individual user model, then the average accuracy scores were reported.

Despite the slight edge in predictive performance by the RFC model, the Stan (Bayesian) model offers significant advantages in terms of interpretability. Unlike its non-probabilistic counterparts, the Stan model provides posterior distributions for the learned weights of molecular descriptors. This feature not only allows for a clear understanding of the importance of different descriptors but also incorporates the uncertainties associated with these weights. Such probabilistic insights are crucial for gaining a deeper understanding of the factors driving experts' preferences and ensuring that the model's predictions are not only accurate but also comprehensible and justifiable.

In summary, while the RFC model boasts the highest predictive accuracy, the Stan model's interpretability and ability to quantify uncertainties make it a valuable tool for elucidating the rationale behind experts' preferences in molecular design. This dual benefit of accuracy and interpretability underscores the potential of Bayesian models in explaining complex decision-making processes such as human rating.

5 Discussion

In this work, we developed and evaluated models to decipher and predict human preferences in molecular design, focusing on the interpretation of these preferences using various, known and self-explanatory molecular descriptors. The Stan (Bayesian) model, Logistic Regression (LogReg), and Random Forest Classifier

(RFC) were employed to fit the preference data provided by three experts on a set of DRD2 binders proposed by a molecular design tool.

The posterior distributions of the Stan model provided insights into the importance of different molecular descriptors for each expert, revealing distinct patterns in their preferences. Expert 1 focused on the structural complexity of molecules and presence of undesired structures for drug-likeness, as evidenced by higher correlations with QED structural alerts. Expert 2's preferences indicated a focus on lipophilicity and synthetic accessibility, similar to Expert 3. Interestingly, all three experts were characterized by high correlations with the DRD2 activity descriptor, aligning with the core objective of the feedback exercise which is to rate DRD2 binders. Notably, the interpretations derived from the Stan model aligned the closest with the reasoning process described by the experts themselves, enhancing the model's ability to accurately explain the expert preference data and decision-making.

The interpretability analysis from the RFC models also highlighted the importance of the DRD2 activity descriptor. For Expert 1, the RFC model did not capture the reliance on structural alerts. For Experts 2 and 3, the molecular LogP and synthetic accessibility descriptors were correctly identified as important, although the RFC model also highlighted molecular weight as a significant factor, which was not explicitly mentioned in expert descriptions. The LogReg models, however, provided a less accurate interpretation.

In terms of predictive accuracy, the RFC model achieved the highest performance, followed by the Stan model and the LogReg model. Despite the superior predictive accuracy of the RFC model, the Stan model's ability to better capture the relationships between the human reasoning processes and the molecular descriptors, and to quantify uncertainties through the posterior distributions, makes it a more interpretable and insightful tool for understanding the reasoning behind experts' preferences.

One of the main limitations of this study is the small amount of expert preference data available. This limited data set may not fully capture the variability and complexity of experts' decision-making processes. Consequently, the generalizability of our models to new, unseen data remains an open question. Future work should focus on collecting more extensive preference data from a larger and more diverse group of experts. This would not only improve the robustness and generalizability of our model but also provide a more comprehensive understanding of how different molecular descriptors influence human preferences in molecular design.

Additionally, it would be valuable to validate the model on unseen data to assess their predictive performance in real-world scenarios. This validation step is crucial for ensuring their practical applicability in molecular design tasks.

In conclusion, while our preliminary model demonstrate high predictive accuracy and provide valuable insights into the reasoning behind experts' preferences, addressing the limitations related to data size and generalizability are essential steps for future work.

Acknowledgments. This study was partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832 "Advanced Machine Learning for Innovative Drug Discovery". Further, this work was supported by the Academy of Finland Flagship program: the Finnish Center for Artificial Intelligence FCAI, and the UKRI Turing AI World-Leading Researcher Fellowship, [EP/W002973/1].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. RDKit: open-source cheminformatics. https://www.rdkit.org
2. Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nat. Chem. **4**(2), 90–98 (2012). https://doi.org/10.1038/NCHEM.1243
3. Blaschke, T., et al.: REINVENT 2.0: an AI tool for de novo drug design. J. Chem. Inf. Model. **60**(12), 5918–5922 (2020)
4. Carpenter, B., et al.: Stan: a probabilistic programming language. J. stat. softw. **76**(1) (2017)
5. Choung, O.H., Vianello, R., Segler, M., Stiefl, N., Jiménez-Luna, J.: Learning chemical intuition from humans in the loop. ChemRxiv (2023). https://doi.org/10.26434/chemrxiv-2023-knwnv
6. Congreve, M., Carr, R., Murray, C., Jhoti, H.: A 'rule of three' for fragment-based lead discovery? Drug discovery today **8**(19), 876–877 (2003). https://doi.org/10.1016/S1359-6446(03)02831-9, https://www.sciencedirect.com/science/article/pii/S1359644603028319
7. Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of cheminformatics **1**, 1–11 (2009)
8. Fennell, P.G., Zuo, Z., Lerman, K.: Predicting and explaining behavioral data with structured feature space decomposition. EPJ Data Sci. **8**(1), 1–27 (2019)
9. Janosch, M., Nahal, Y., Jannik Bjerrum, E., Kabeshov, M., Engkvist, O., Kaski, S.: A python-based user interface to collect expert feedback for generative chemistry models. ChemRxiv. 2024; https://doi.org/10.26434/chemrxiv-2024-zs5xp This content is a preprint and has not been peer-reviewed (2024)
10. Kutchukian, P.S., et al.: Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. PLoS ONE **7**(11), e48476 (2012)
11. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017). https://arxiv.org/abs/1705.07874
12. Maimon, O., Rokach, L.: Improving supervised learning by feature decomposition. In: Eiter, T., Schewe, K.D. (eds.) Foundations of Information and Knowledge Systems, pp. 178–196. Springer, Berlin Heidelberg, Berlin, Heidelberg (2002). https://doi.org/10.1007/3-540-45758-5_12
13. Olivecrona, M., Blaschke, T., Engkvist, O., Chen, H.: Molecular de-novo design through deep reinforcement learning. J. Cheminformatics **9**(1), 48 (2017). https://doi.org/10.1186/s13321-017-0235-x

14. Pedregosa, F., et al.: SciKit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
15. Sundin, I., et al.: Human-in-the-loop assisted de novo molecular design. J. Cheminformatics **14**(1), 86 (2022). https://doi.org/10.1186/s13321-022-00667-8