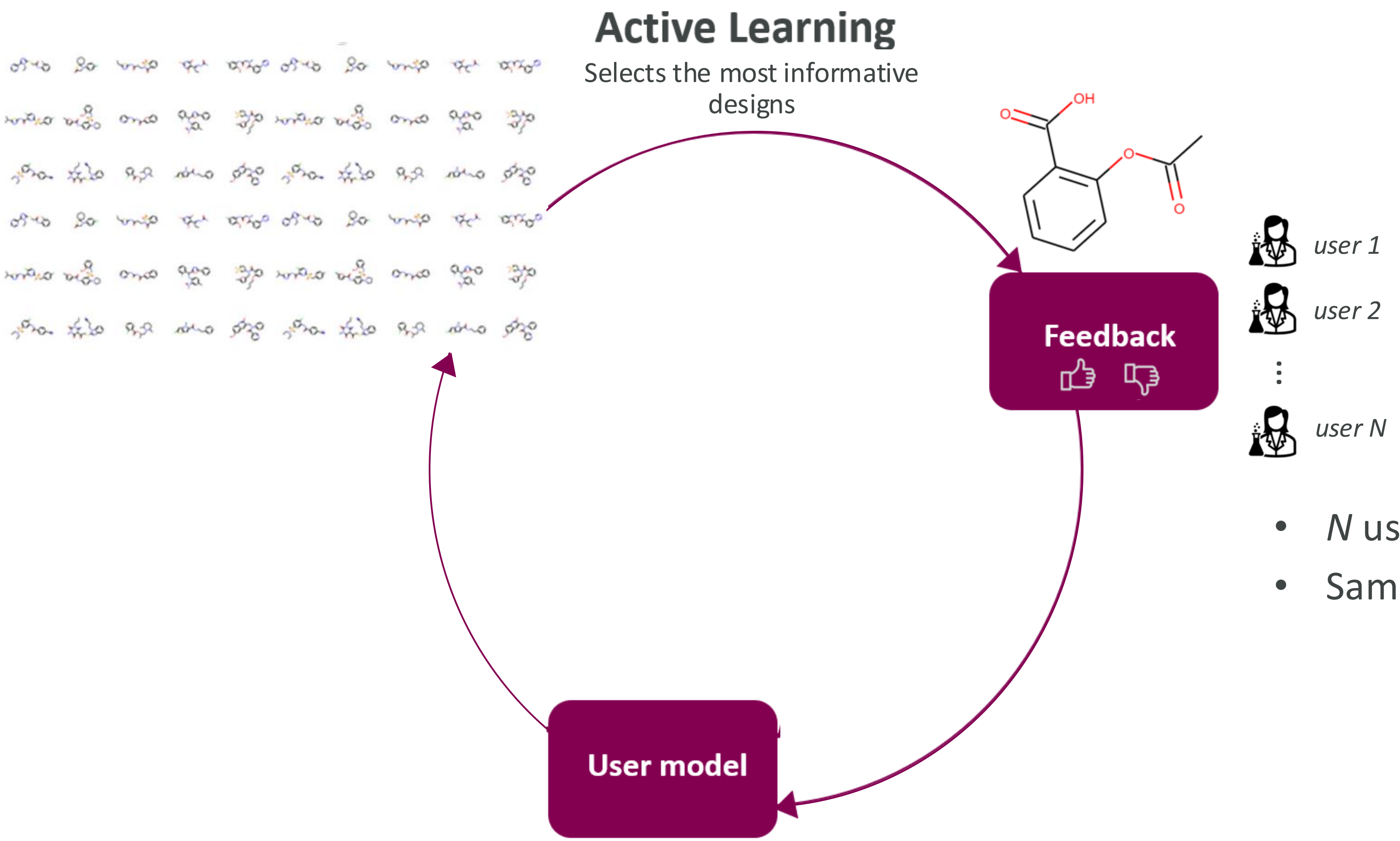
Towards Interpretable Models of Chemist Preferences for *De novo* Molecular Design

ICANN 2024 – AI in Drug Discovery Workshop

Yasmine Nahal

Motivation



Active Learning

Selects the most informative designs

Feedback

User model

user 1

user 2

⋮

user N

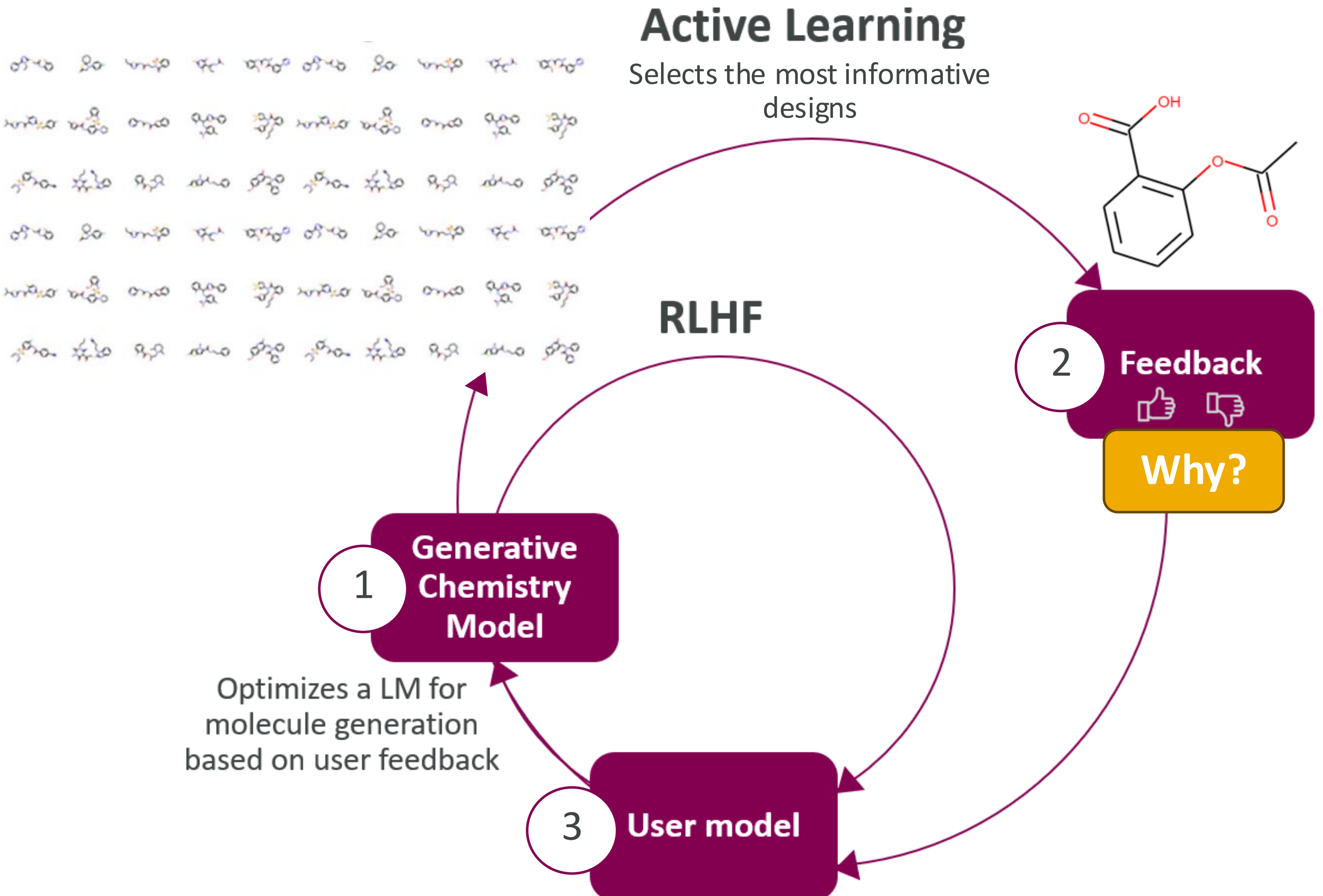
- *N* users
- Same design goal

Motivation

I would rather use metrics that ***I understand*** like the QED score

How can we make HITL ML for drug design more practical for the community?

Build **interpretable models of chemist preferences** that can effectively integrate into drug design workflows



Active Learning

Selects the most informative designs

RLHF

1 **Generative Chemistry Model**

Optimizes a LM for molecule generation based on user feedback

2 **Feedback**

Why?

3 **User model**

user 1

user 2

⋮

user N

- *N* users
- Same design goal

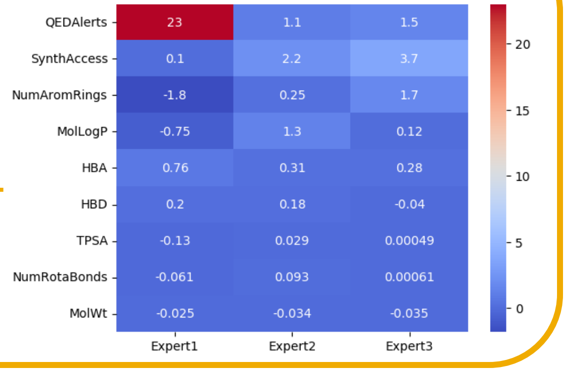
Feature Selection

user 1

user 2

⋮

user N

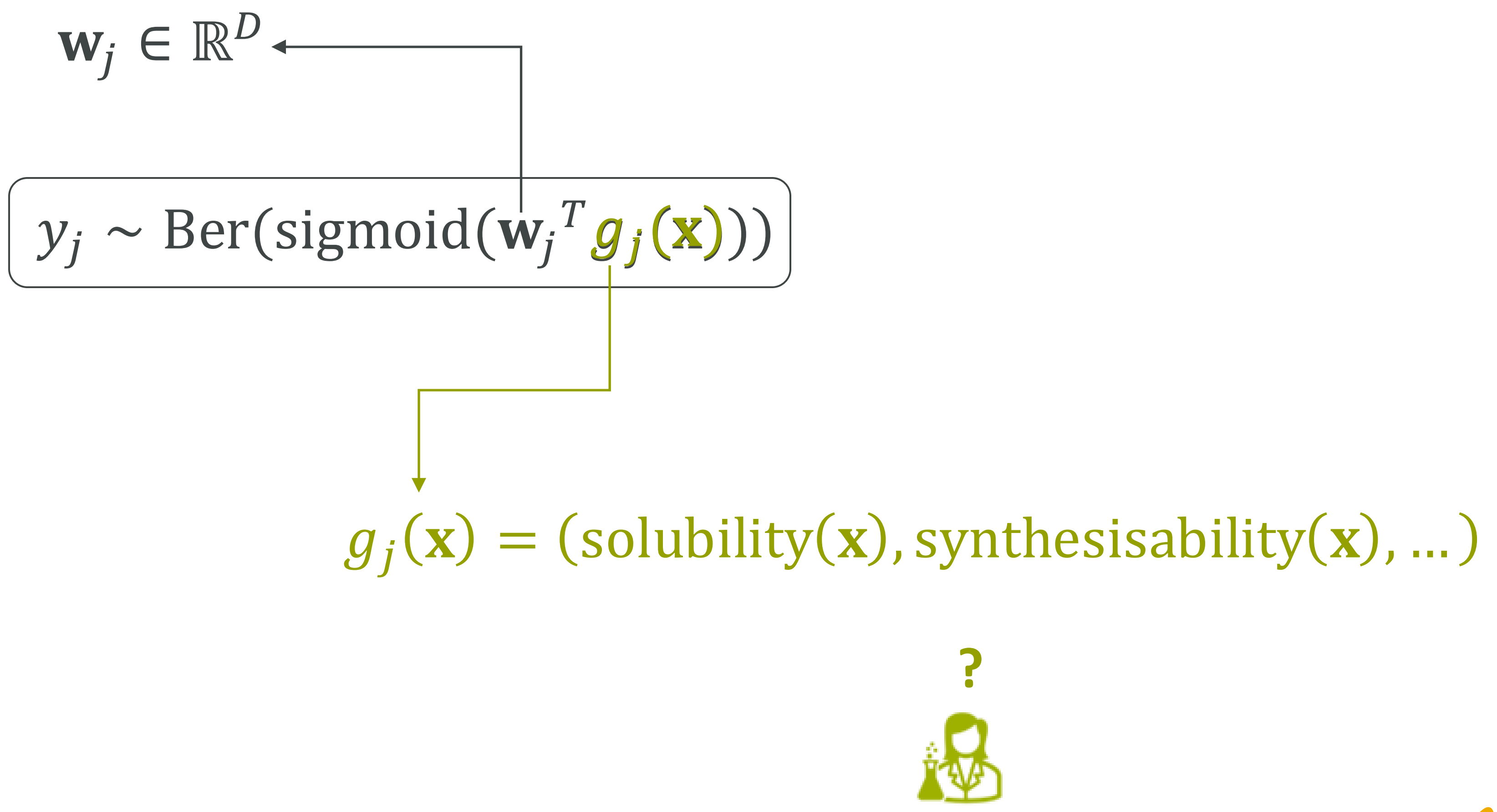


	Expert1	Expert2	Expert3
QEDAlerts	23	1.1	1.5
SynthAccess	0.1	2.2	3.7
NumAromRings	-1.8	0.25	1.7
MolLogP	-0.75	1.3	0.12
HBA	0.76	0.31	0.28
HBD	0.2	0.18	-0.04
TPSA	-0.13	0.029	0.00049
NumRotaBonds	-0.061	0.093	0.00061
MolWt	-0.025	-0.034	-0.035

Methodology

Setting

We consider J user responses about $\mathbf{x}$, $Y_j = \{(\mathbf{x}_{ij}, y_{ij})\}_{i=1}^{N}$ where



$\mathbf{w}_j \in \mathbb{R}^D$

$y_j \sim \mathrm{Ber}(\mathrm{sigmoid}({\mathbf{w}_j}^T g_j(\mathbf{x})))$

$g_j(\mathbf{x}) = (\mathrm{solubility}(\mathbf{x}), \mathrm{synthesisability}(\mathbf{x}), \dots)$

Setting

We assume that all users share the same g with different weights $\mathbf{w}_j$

→ *The set of features used by any expert is the union of all features*

- Likelihood

$$\boxed{y_j \sim \mathrm{Ber}(\mathrm{sigmoid}({\mathbf{w}_j}^T g(\mathbf{x})))}$$

Setting

We assume that all users share the same g with different weights $\mathbf{w}_j$

→ *The set of features used by any expert is the union of all features*

- Likelihood

$$\boxed{y_j \sim \mathrm{Ber}(\mathrm{sigmoid}(\mathbf{w}_j{}^T g(\mathbf{x})))}$$

→ *Unused features by an expert will show as zeros in $\mathbf{w}_j$*

- Sparse prior

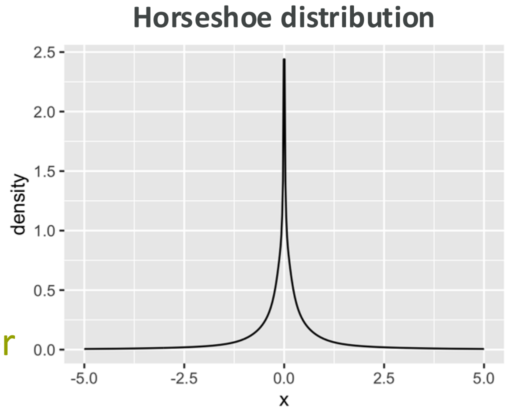
$$\boxed{p(\mathbf{w}_j) = HS(\mathbf{w}_j) = N(0, \boldsymbol{\lambda}_j^2 . \boldsymbol{\tau}^2)}$$

Global scale parameter

Local scale parameter for each user j

$$p(\boldsymbol{\lambda}_j) = \ Cauchy(\boldsymbol{\lambda}_j)$$

$$p(\boldsymbol{\tau}) = \ Cauchy(\boldsymbol{\tau})$$



Setting

We assume that all users share the same g with different weights $\mathbf{w}_j$

→ The set of features used by any expert is the union of all features

- Likelihood

$$y_j \sim \mathrm{Ber}(\mathrm{sigmoid}(\mathbf{w}_j{}^T g(\mathbf{x})))$$

→ Unused features by an expert will show as zeros in $\mathbf{w}_j$

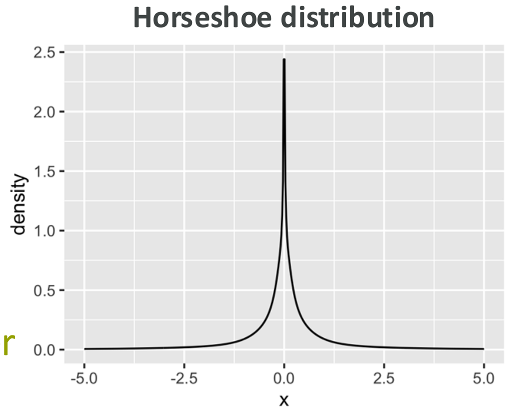
- Sparse prior

$$p(\mathbf{w}_j) = HS(\mathbf{w}_j) = N(0, \boldsymbol{\lambda}_j^2 . \boldsymbol{\tau}^2)$$

Global scale parameter

Local scale parameter for each user j

$$p(\boldsymbol{\lambda}_j) = Cauchy(\boldsymbol{\lambda}_j)$$
$$p(\boldsymbol{\tau}) = Cauchy(\boldsymbol{\tau})$$


Horseshoe distribution

- Posterior

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

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_J) \in \mathbb{R}^{J \times D}$

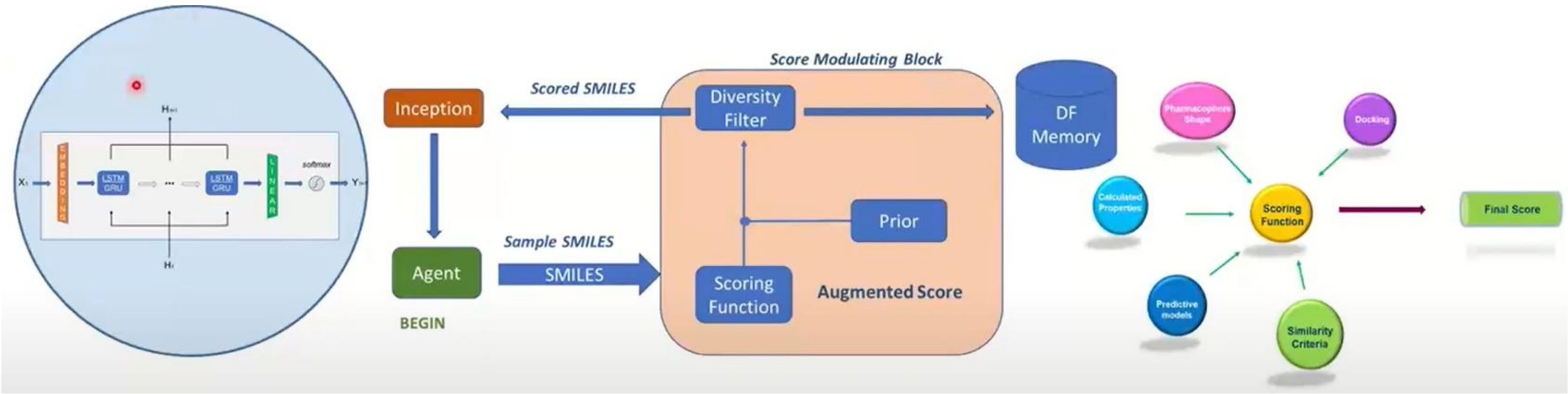
- ✓ **Interpretable features**
- ✓ **Built-in uncertainty**

Initial experiments

De novo molecular design

- **Step 1:** design novel DRD2 binders

REINVENT



Generative Seq2Seq Model

Reinforcement Learning

Composite scoring function

- **QED score**
- **hERG-QSAR**
- **DRD2-QSAR**

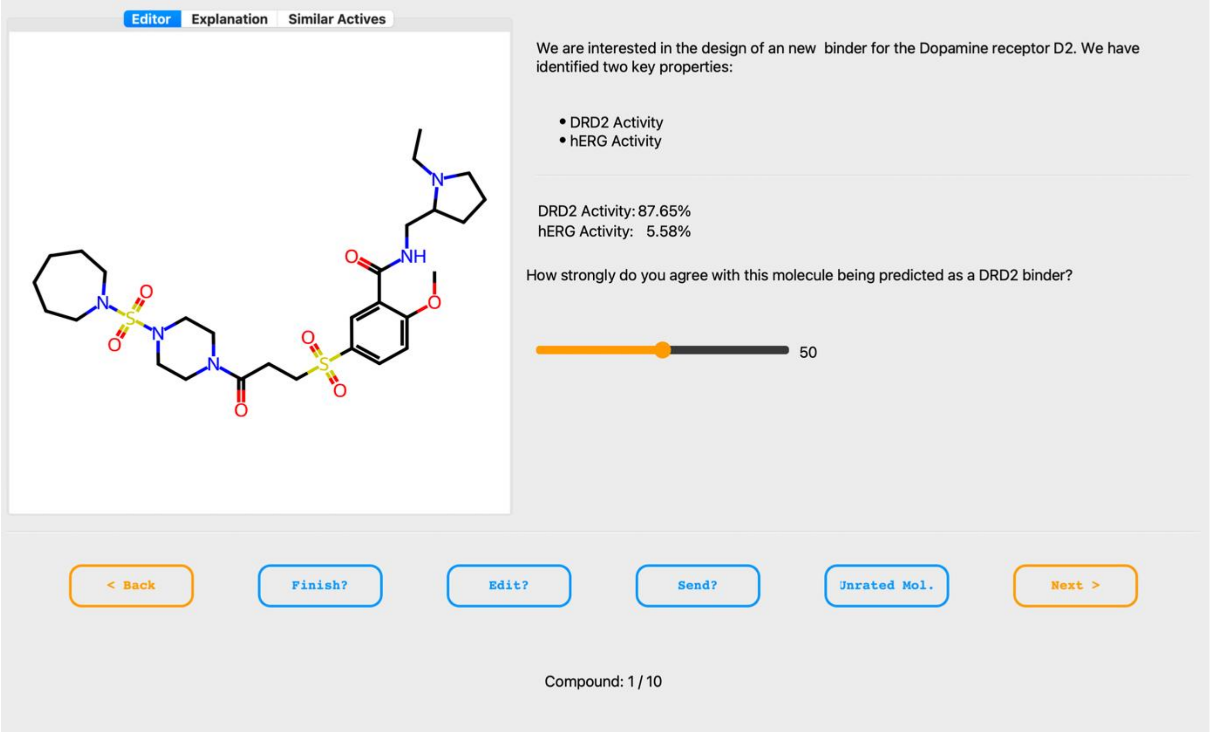
Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., Papadopoulos, K., & Patronov, A. (2020). REINVENT 2.0: An AI Tool for De Novo Drug Design. *Journal of chemical information and modeling*.

De novo molecular design with user feedback

- **Step 2:** select a set of high-scored DRD2 binders to be labelled by the user

- **Step 3:** fine-tune the DRD2-QSAR model with user feedback

- **Step 4:** resume the design process using the refined scoring model ***(RLHF)***



We are interested in the design of an new binder for the Dopamine receptor D2. We have identified two key properties:

- DRD2 Activity
- hERG Activity

DRD2 Activity: 87.65%
hERG Activity: 5.58%

How strongly do you agree with this molecule being predicted as a DRD2 binder?

50

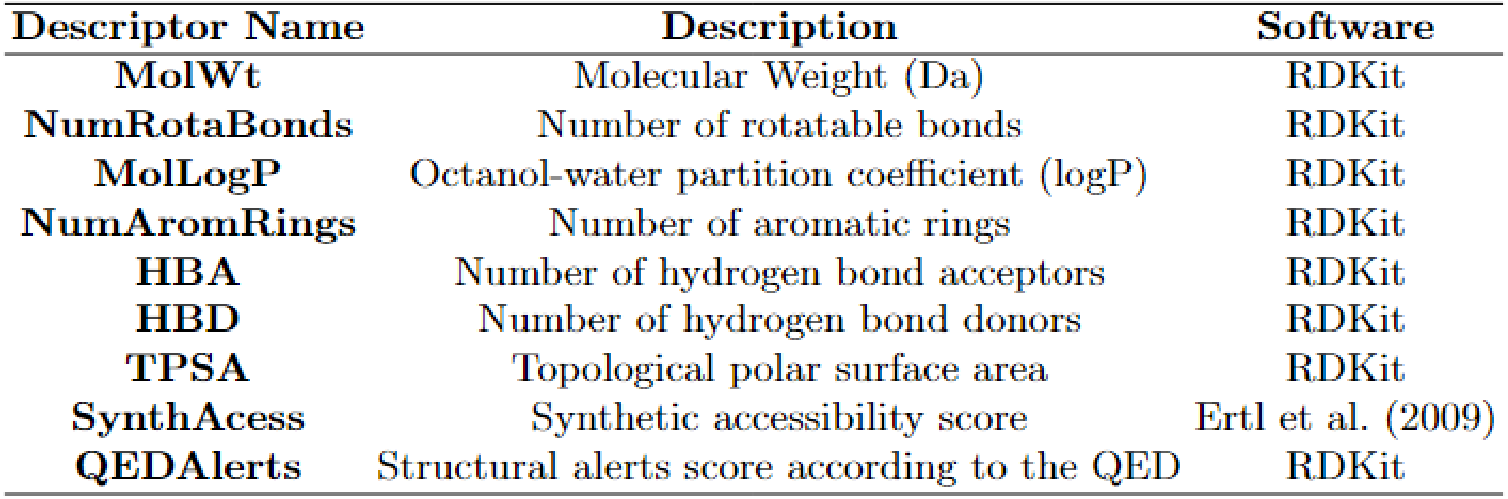
Compound: 1 / 10

- 3 expert participants from AstraZeneca
- 150 actively selected molecules labelled by each expert

Nahal Y, Menke J, Martinelli J, Heinonen M, Kabeshov M, Janet JP, et al. Human-in-the-loop active learning for goal-oriented molecule generation. ChemRxiv. 2024.
Menke, J., Nahal, Y., Bjerrum, E.J. *et al.* `Metis`: a python-based user interface to collect expert feedback for generative chemistry models. *J Cheminform* **16**, 100 (2024).

Molecular features

- 2D physchem descriptors
- 2048 ECFP6



Descriptor Name	Description	Software
MolWt	Molecular Weight (Da)	RDKit
NumRotaBonds	Number of rotatable bonds	RDKit
MolLogP	Octanol-water partition coefficient (logP)	RDKit
NumAromRings	Number of aromatic rings	RDKit
HBA	Number of hydrogen bond acceptors	RDKit
HBD	Number of hydrogen bond donors	RDKit
TPSA	Topological polar surface area	RDKit
SynthAcess	Synthetic accessibility score	Ertl et al. (2009)
QEDAlerts	Structural alerts score according to the QED	RDKit

Feature selection

Posterior inference

- Stan programming language
- MCMC sampling

  (2 chains, 2000 iterations)

```
data {
  int<lower=0> N;       // number of molecules
  int<lower=0> J;       // number of experts
  int<lower=0> D;       // number of molecular descriptors
  matrix[N, D] X;       // molecular descriptors
  int<lower=0, upper=1> Y[N, J];  // binary responses from experts
  real<lower=0> tau_0;  // global shrinkage parameter
}

parameters {
  real<lower=0> tau;           // global scale parameter
  vector<lower=0>[D] lam[J];   // local scale parameters
  matrix[D, J] w;              // preference weights
}

model {
  // Horseshoe prior
  tau ~ cauchy(0, tau_0);
  for (j in 1:J) {
    lam[j] ~ cauchy(0, 1);
    for (d in 1:D) {
      w[d, j] ~ normal(0, lam[j][d] * tau);
    }
  }

  // Likelihood
  for (n in 1:N) {
    for (j in 1:J) {
  //Y ~ bernoulli_logit(X * w);
      Y[n, j] ~ bernoulli_logit(dot_product(w[, j], X[n, ]));
    }
  }
}
```

Benchmark

Feature selection methods

- LASSO Logistic Regression
- Sparse Neural Network Classifier *(3 hidden layers, softmax output)*
- Random Forest Classifier

Expert descriptions of their reasonings at the end of the process

Expert 1: “I looked at the structures of known DRD2 actives to judge if the new designed ones are relevant. I disliked molecules that contained undesirable substructures.”

Expert 2: “I assessed how much I liked the molecule as a lead, so I selected molecules that would be synthesisable, stable and with reasonable lipophilicity to give them the best chance for being made and tested in a project. No prior experience with the SAR.”

Expert 3: “I didn’t have much knowledge about the DRD2 target. I selected molecules that synthetic chemists would be willing to test. ”

Performance metrics

- Predictive accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- User agreement

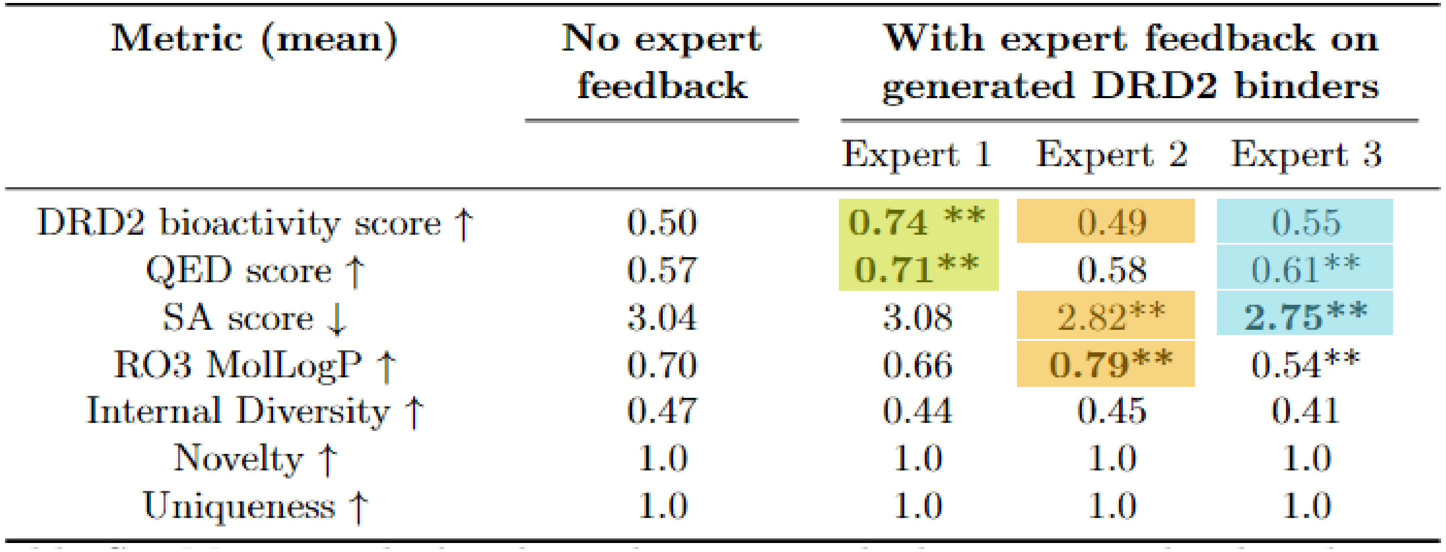
Results

Expert feedback improves *de novo* molecular design

At the end of the design process, we **selected the final set of high-scored DRD2 binders**.

- *Is the design goal achieved after introducing expert feedback ?*

- *How right was each expert about their reasoning in comparison with no feedback ?*



Metric (mean)	No expert feedback	With expert feedback on generated DRD2 binders		
		Expert 1	Expert 2	Expert 3
DRD2 bioactivity score ↑	0.50	**0.74 ****	0.49	0.55
QED score ↑	0.57	**0.71****	0.58	0.61**
SA score ↓	3.04	3.08	2.82**	**2.75****
RO3 MolLogP ↑	0.70	0.66	**0.79****	0.54**
Internal Diversity ↑	0.47	0.44	0.45	0.41
Novelty ↑	1.0	1.0	1.0	1.0
Uniqueness ↑	1.0	1.0	1.0	1.0

Expert 1: “I looked at the structures of known DRD2 actives to judge if the new designed ones are relevant. I disliked molecules that contained undesirable substructures.”

Expert 2: “I assessed how much I liked the molecule as a lead, so I selected molecules that would be synthesisable, stable and with reasonable lipophilicity to give them the best chance for being made and tested in a project. No prior experience with the SAR.”

Expert 3: “I didn’t have much knowledge about the DRD2 target. I selected molecules that synthetic chemists would be willing to test.”

Nahal Y, Menke J, Martinelli J, Heinonen M, Kabeshov M, Janet JP, et al. Human-in-the-loop active learning for goal-oriented molecule generation. ChemRxiv. 2024.

Bayesian feature selection performs equally or better than non-Bayesian alternatives

(a) Models trained on 2D molecular descriptors

	LASSO LogReg (L1 regularization)	Sparse NN (L1 regularization)	Random Forest	Bayesian LogReg (sparse prior)
Mean Train Accuracy	0.81	0.85	0.99	0.89
Mean Test accuracy (Stratified 80/20 split)	0.69	0.81	0.82	0.85

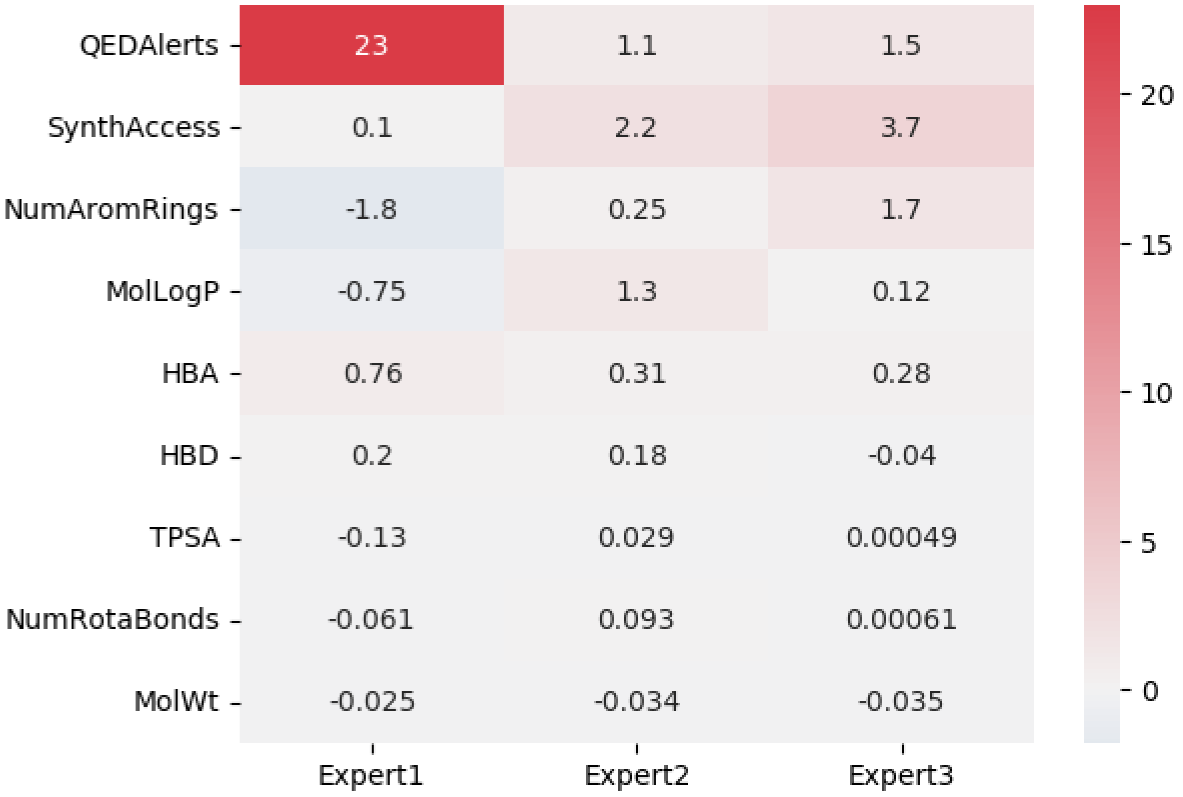
(b) Models trained on 2D molecular descriptors + ECFPs

	LASSO LogReg (L1 regularization)	Sparse NN (L1 regularization)	Random Forest	Bayesian LogReg (sparse prior)
Mean Train Accuracy	0.86	0.91	0.99	0.96
Mean Test Accuracy (Stratified 80/20 split)	0.70	0.78	0.85	0.83

Bayesian feature selection aligns well with expert descriptions

(a) Models trained on 2D molecular descriptors

Descriptor mean weight magnitude extracted from the learned posterior distribution of the weights



	Expert1	Expert2	Expert3
QEDAlerts	23	1.1	1.5
SynthAccess	0.1	2.2	3.7
NumAromRings	-1.8	0.25	1.7
MolLogP	-0.75	1.3	0.12
HBA	0.76	0.31	0.28
HBD	0.2	0.18	-0.04
TPSA	-0.13	0.029	0.00049
NumRotaBonds	-0.061	0.093	0.00061
MolWt	-0.025	-0.034	-0.035

Expert 1: “I looked at the structures of known DRD2 actives to judge if the new designed ones are relevant. I disliked molecules that contained undesirable substructures.”

Expert 2: “I assessed how much I liked the molecule as a lead, so I selected molecules that would be synthesisable, stable and with reasonable lipophilicity to give them the best chance for being made and tested in a project. No prior experience with the SAR.”

Expert 3: “I didn’t have much knowledge about the DRD2 target. I selected molecules that synthetic chemists would be willing to test. ”

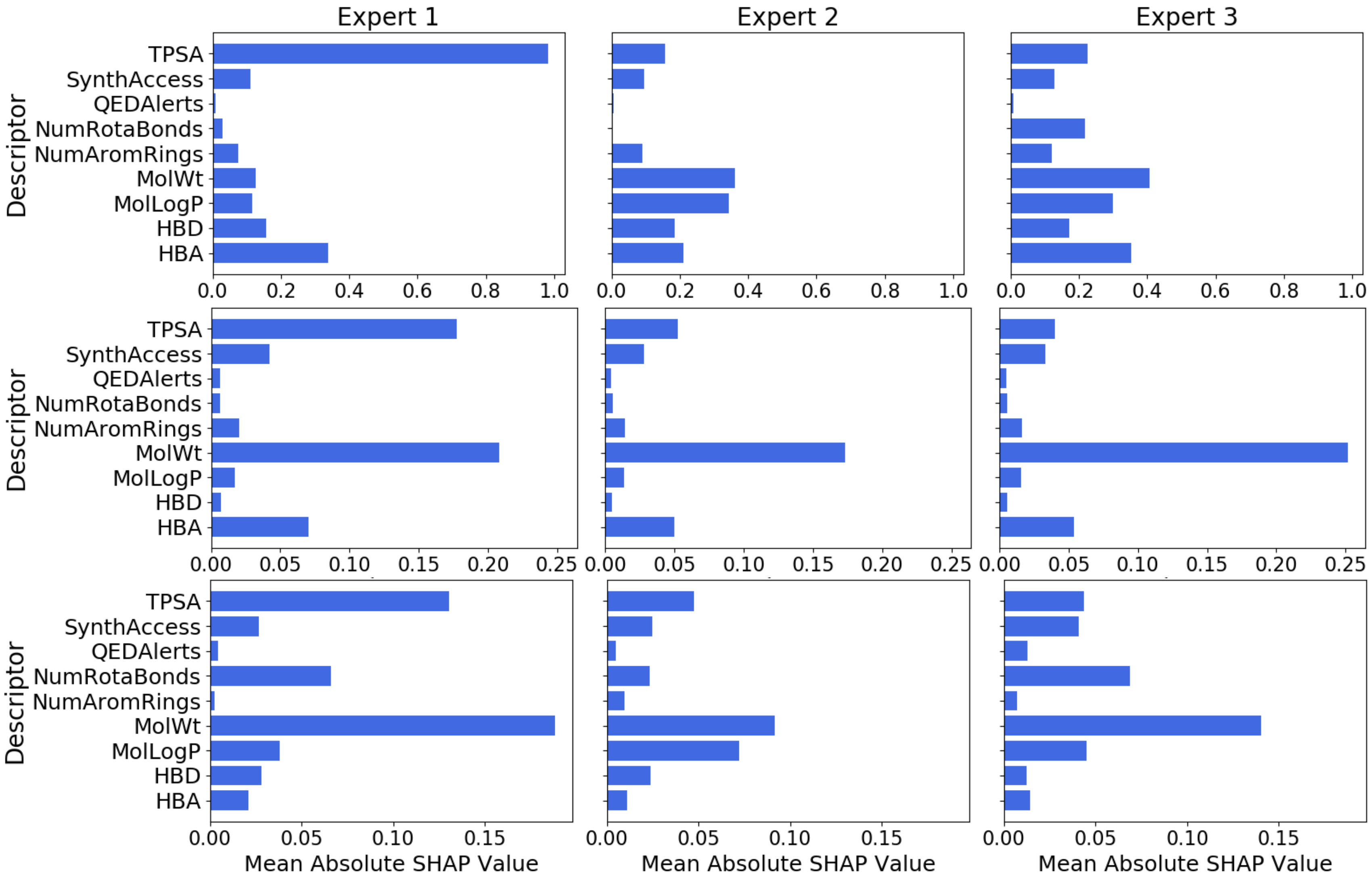
Alternative methods align less with expert descriptions

(a) Models trained on 2D molecular descriptors

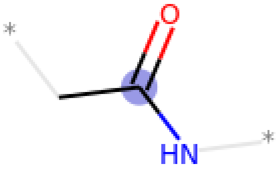
LASSO LogReg

Sparse NN

Random Forest

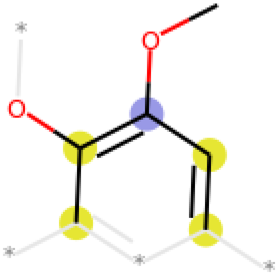


(b) Models trained on 2D molecular descriptors + ECFPs



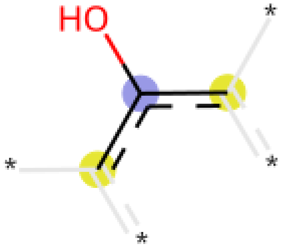
bit41

All molecules containing this motif were disliked by Expert 1



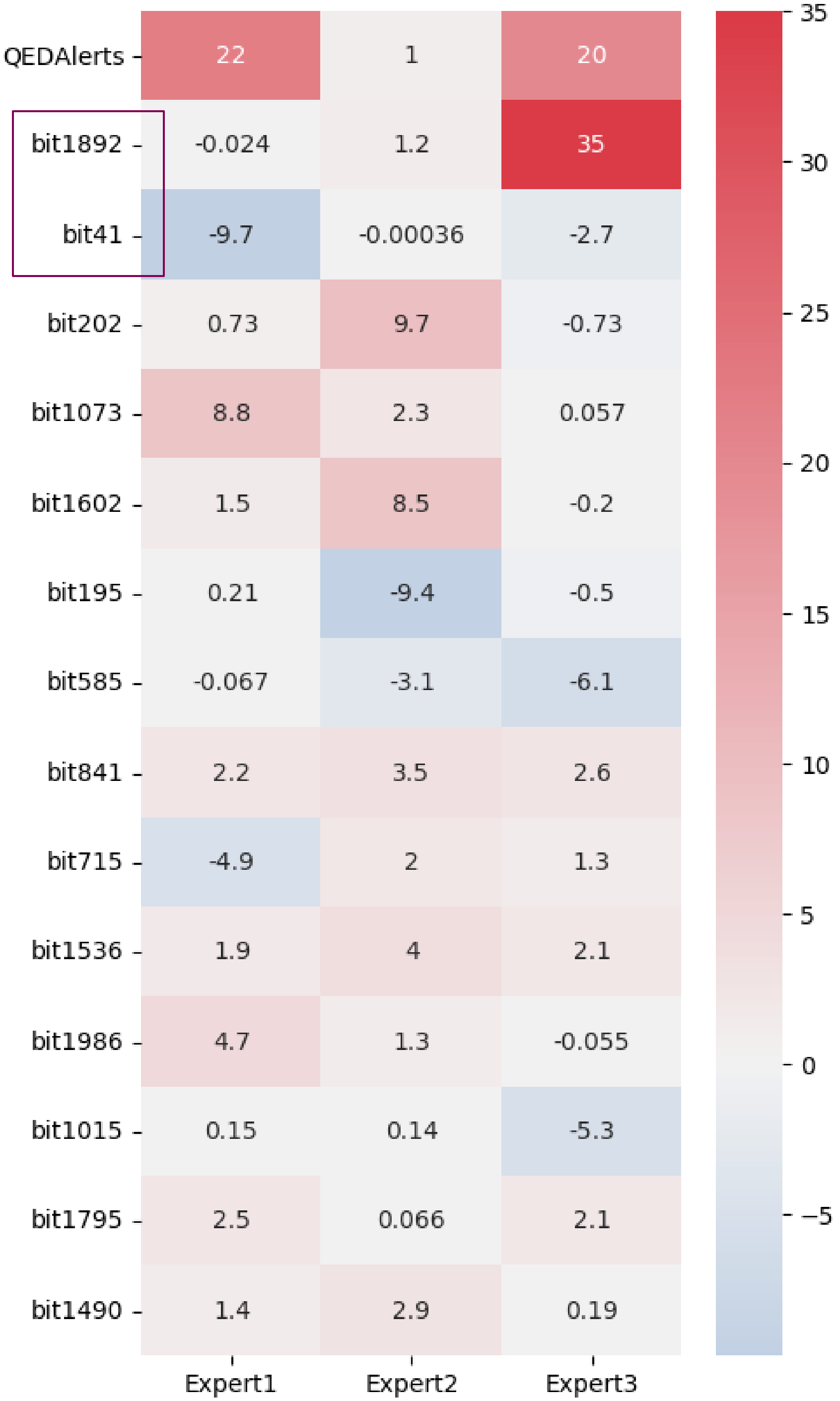
bit1892

All molecules containing this motif were liked by Expert 3



bit202

All molecules containing this motif were liked by Expert 2



	Expert1	Expert2	Expert3
QEDAlerts	22	1	20
bit1892	-0.024	1.2	35
bit41	-9.7	-0.00036	-2.7
bit202	0.73	9.7	-0.73
bit1073	8.8	2.3	0.057
bit1602	1.5	8.5	-0.2
bit195	0.21	-9.4	-0.5
bit585	-0.067	-3.1	-6.1
bit841	2.2	3.5	2.6
bit715	-4.9	2	1.3
bit1536	1.9	4	2.1
bit1986	4.7	1.3	-0.055
bit1015	0.15	0.14	-5.3
bit1795	2.5	0.066	2.1
bit1490	1.4	2.9	0.19

Expert 1: "I looked at the structures of known DRD2 actives to judge if the new designed ones are relevant. I disliked molecules that contained undesirable substructures."

Expert 2: "I assessed how much I liked the molecule as a lead, so I selected molecules that would be synthesisable, stable and with reasonable lipophilicity to give them the best chance for being made and tested in a project. No prior experience with the SAR."

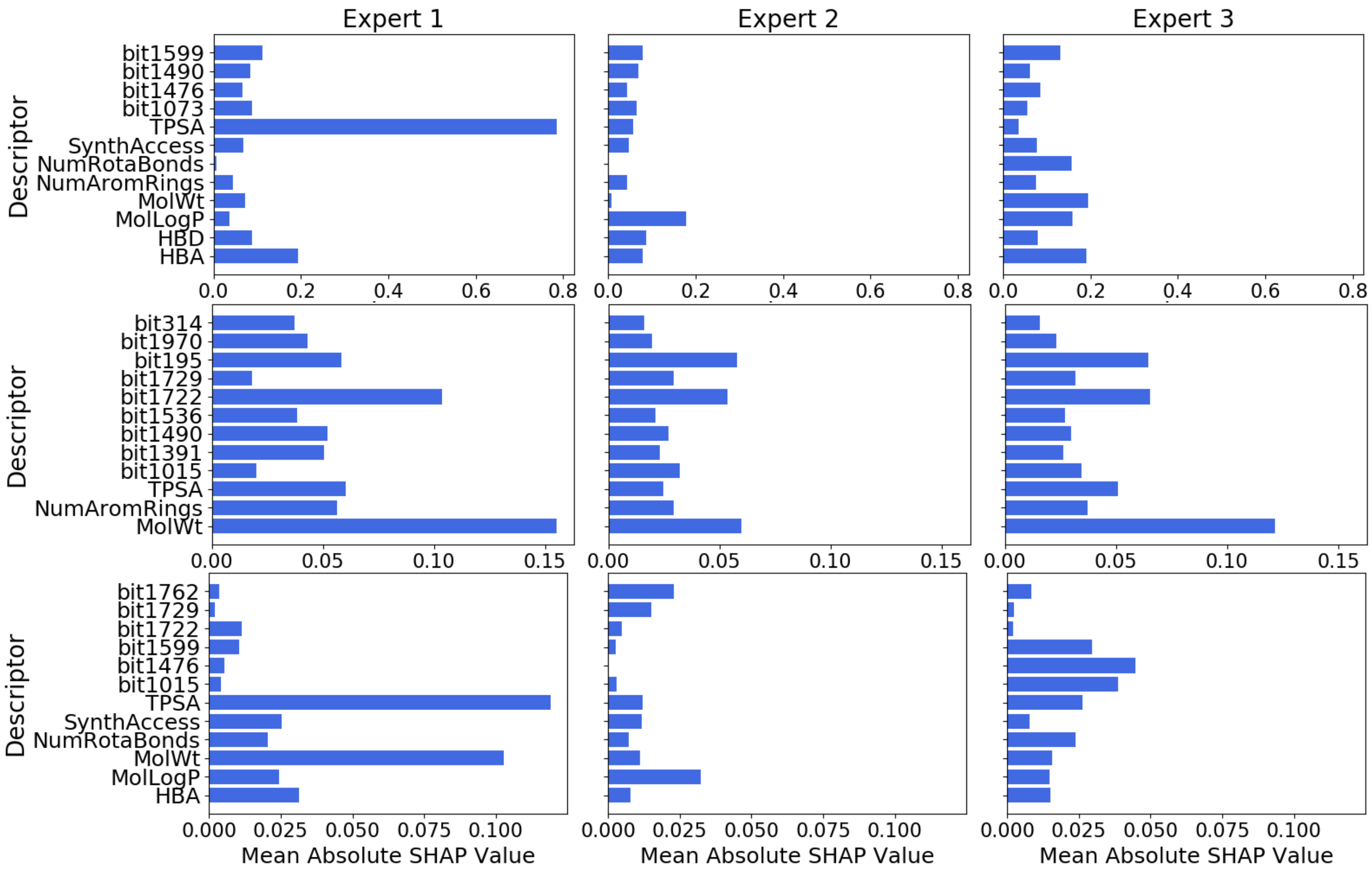
Expert 3: "I didn't have much knowledge about the DRD2 target. I selected molecules that synthetic chemists would be willing to test. "

(b) Models trained on 2D molecular descriptors + ECFPs

LASSO LogReg

Sparse NN

Random Forest

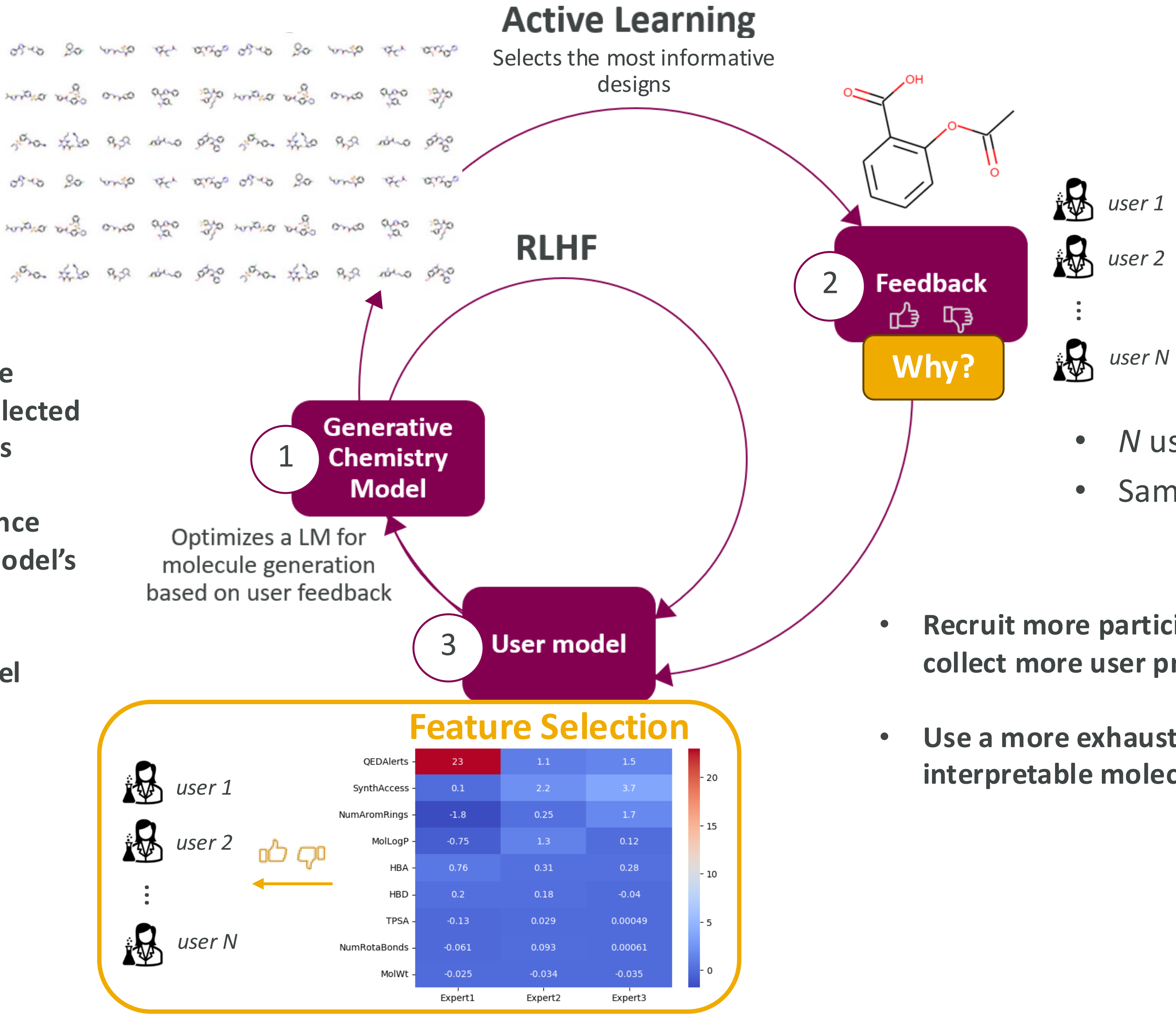


Summary

- We aim to enhance the transparency and practical usability of user models in drug design.
- Our method integrates Bayesian inference with a sparse prior to build interpretable chemistry user models.
- The Bayesian method outperforms the Lasso logistic regression and the sparse neural network in predicting user responses.
- The Bayesian method’s interpretable feature importances are the closest to user-written descriptions.

Future work: enhanced alignment with expert reasoning

- Users provide feedback on the importance of the features selected in explaining their preferences
- Feedback on feature importance directly influences the user model's learning process
- The feedback will adjust model predictions to reflect which features align with expert reasoning


Active Learning
Selects the most informative designs
RLHF
1
Generative Chemistry Model
Optimizes a LM for molecule generation based on user feedback
2
Feedback
Why?
3
User model
user 1
user 2
user N
Feature Selection
user 1
user 2
user N
O
OH
O
O
20
15
10
5
0

	Expert1	Expert2	Expert3
QEDAlerts	23	1.1	1.5
SynthAccess	0.1	2.2	3.7
NumAromRings	-1.8	0.25	1.7
MolLogP	-0.75	1.3	0.12
HBA	0.76	0.31	0.28
HBD	0.2	0.18	-0.04
TPSA	-0.13	0.029	0.00049
NumRotaBonds	-0.061	0.093	0.00061
MolWt	-0.025	-0.034	-0.035

- *N* users
- Same design goal

- Recruit more participants and collect more user preference data
- Use a more exhaustive list of interpretable molecular features

Thank you for your attention!

Future work: enhanced alignment with expert reasoning

The user can give feedback about the selected features m

- Prior over their weights: $w_{j,m} \sim \gamma_{j,m}\, N\left(0, \lambda_{j,m}^2\right) + \left(1 - \gamma_{j,m}\right)\delta_0,$

$$\text{where } \gamma_{j,m} \sim \mathrm{Ber}\left(\rho_j\right),$$

$$\text{and } \rho_j \sim \mathrm{Beta}(\alpha_j^{\rho}, \beta_j^{\rho})$$

- User feedback a feature importance:

$$z_{j,m} \sim \gamma_{j,m}\mathrm{Ber}\left(\pi_j\right) + \left(1 - \gamma_{j,m}\right)\mathrm{Ber}(1 - \pi_j)$$

$$\text{where } \pi_j \sim \mathrm{Beta}(\alpha_j^{\pi}, \beta_j^{\pi})$$

- Joint posterior:

$$p\left(\boldsymbol{\theta_j} \,\middle|\, Y_j, Z_j\right) \propto \prod_j p\left(Y_j \,\middle|\, \mathbf{w}_j\right) p(Z_j \mid \boldsymbol{\gamma}_j, \boldsymbol{\pi}_j)\; p(\mathbf{w}_j \mid \boldsymbol{\lambda_j^2}, \boldsymbol{\gamma_j})\; p(\boldsymbol{\gamma_j} \mid \boldsymbol{\rho_j})\; p(\boldsymbol{\rho_j}) p(\boldsymbol{\pi}_j)$$

$$\text{where } \boldsymbol{\theta_j} = \{\mathbf{w_j}, \boldsymbol{\lambda_j^2}, \boldsymbol{\gamma_j}, \boldsymbol{\rho_j}, \boldsymbol{\pi_j}\}$$

Iiris Sundin, Tomi Peltola et al., Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge, *Bioinformatics*, Volume 34, Issue 13, July 2018, Pages i395-i403, https://doi.org/10.1093/bioinformatics/bty257

Confidentiality Notice