Exploration and validation of large language models as tools for molecular optimization

Christoph Grebner[1], Alejandro Corrochano-Navarro[2], Christian Buning[1], Hans Matter[1], María Méndez[1], Sven Ruf[1], Jan H. Griwatz[1], Elisabeth Speckmeier[1], Thorsten Sadowski[1], Jiří Vymětal[2#], Saeed Moayedpour[2], Lorenzo Kogler-Anele[2], Ziv Bar-Joseph[2], Sven Jager[2*], Gerhard Hessler[1*]

[1]Integrated Drug Discovery, Sanofi, 65926 Frankfurt am Main, Germany.
[2]Digital R&D, Sanofi, Cambridge, MA, USA.
[#]DataSentics, Brno, Czech Republic.

*Corresponding author:
sven.jager@sanofi.com,
gerhard.hessler@sanofi.com

Appendix

Technical details

Our workflow utilizes the LiteLLM package, which allows for seamless usage of a wide variety of deployed LLMs through Bedrock, Hugging Face, OpenAI, etc., as the molecular generation engine. In order to perform many-shot in-context learning (ICL), we used the Claude 3.5 Sonnet model with a 200k context window.

Evaluating the ability of an LLM to learn Structure-Activity relationships from data

To demonstrate the capacity of a Large Language Models to infer structure-activity relationships directly from molecular examples, we designed two complementary tasks: regression and classification of biological activity in what we referred to as SAR-data, that is, compounds from the biaryl series. In parallel, we trained multiple conventional machine learning architectures on three molecular descriptors sets (RDKit, mol2vec, and circular fingerprints), with models trained separately for each descriptor type. These provide a structured representation of molecular features for capturing chemical similarity and SAR-relevant patterns.

Fig. S1 shows that, while the LLM performs competitively and captured SAR trends in a few-shot fashion, descriptor-based ML models consistently achieved slightly higher performance. This gap likely originates from fundamental differences in how the data is processed, as the LLM relies on text-based input in the form of SMILES embedded in a written prompt. The inference must be done with no access to enriched features or lack of direct optimization for the target metric, i.e., fine-

tuning on the task, which may result in worse recognition of subtle physicochemical trends. However, despite these limitations, we do observe promising zero- or few-shot prediction capabilities, presenting a compelling alternative in scenarios where computational resources are limited or curated descriptors are not available.

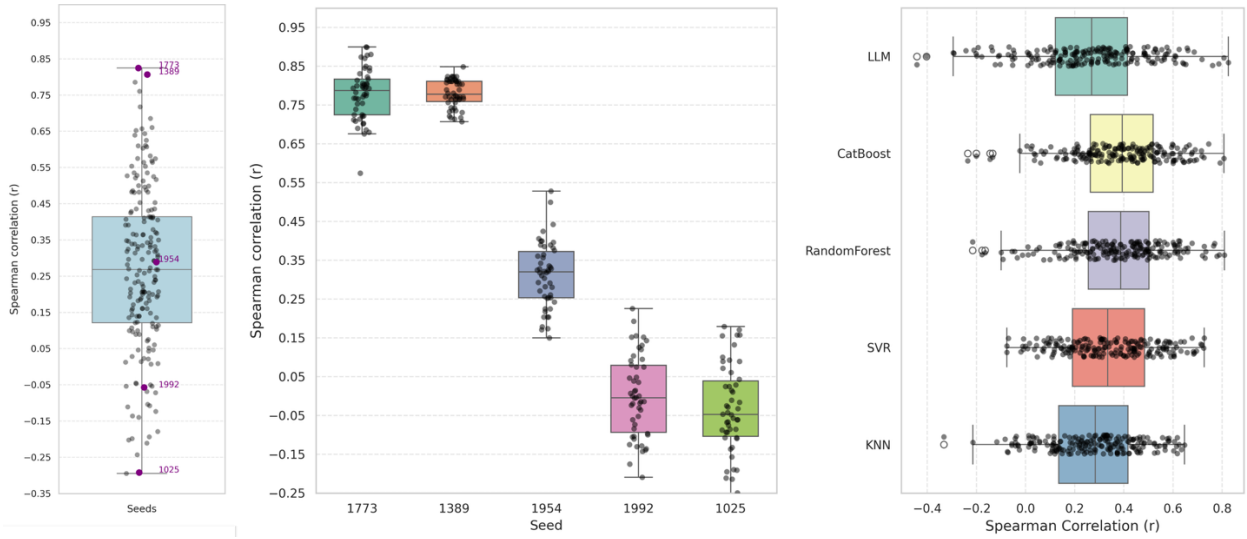


Fig. S1: Performance of LLM-based prediction and descriptor-based machine learning models. The experiment is repeated 200 times for different random splits, where 95% of data is used for training and 5% is used for testing.

In addition to performance, we observed considerable variability across random seeds. This spread is consistent across all model types and descriptors sets, which could be attributable in part to the limited dataset size (296 biaryl) and the small test set employed (95/5 splits). To prove this further, we selected seeds that yielded particularly high or low performance and re-ran each for 50 independent LLM executions. The results revealed consistent trends: certain data splits led to stronger or weaker predictive outcomes, suggesting that variation is not due to LLM's stochasticity, but is closely tied to the reduced number of samples.

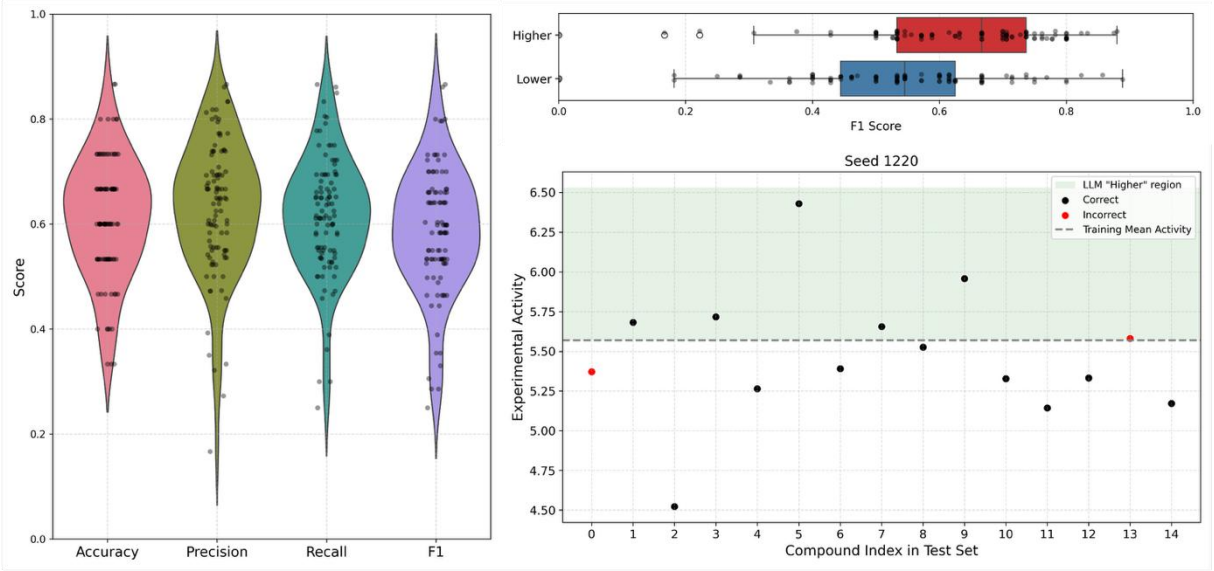


Fig. S2: Classification experiment with LLM models trained on same dataset splits as for regression models above. The experiment is repeated 200 times for different random splits, where 95% of data is used for training and 5% is used for testing.

Fig. S2 provides a broader view of the LLM's ability to distinguish compounds with differing biological activity. We observe LLM's capacity to perform this task, especially when it comes to identifying compounds with higher activity, achieving an average F1-score of 0.64 compared to 0.54 for the lower class. This could suggest the LLM is more confident when recognizing patterns associated with active compounds.

Pairwise classification task:

15 most active and 15 least active molecules were selected. All pairs between active and inactive molecules were generated. The LLM got the task to predict the more active molecules. We first performed this task without any additional data. We then repeated that task with SAR-data selected from the un-selected molecules. For comparison, we evaluated four datasets: the two internal sets (CatA and RIPK1) as well as two public sets (MMP8 and GSK3β[1]). For each target, we added 1000 random molecules as SAR-data. All experiments were performed 3 times.

[1] Supporting information of: S. Sauer, H. Matter, G. Hessler and C. Grebner, "Integrating Reaction Schemes, Reagent Databases, and Virtual Libraries into Fragment-Based Design by Reinforcement Learning," *Journal of Chemical Information and Modeling,* vol. 63, pp. 5709-5726, 2023.

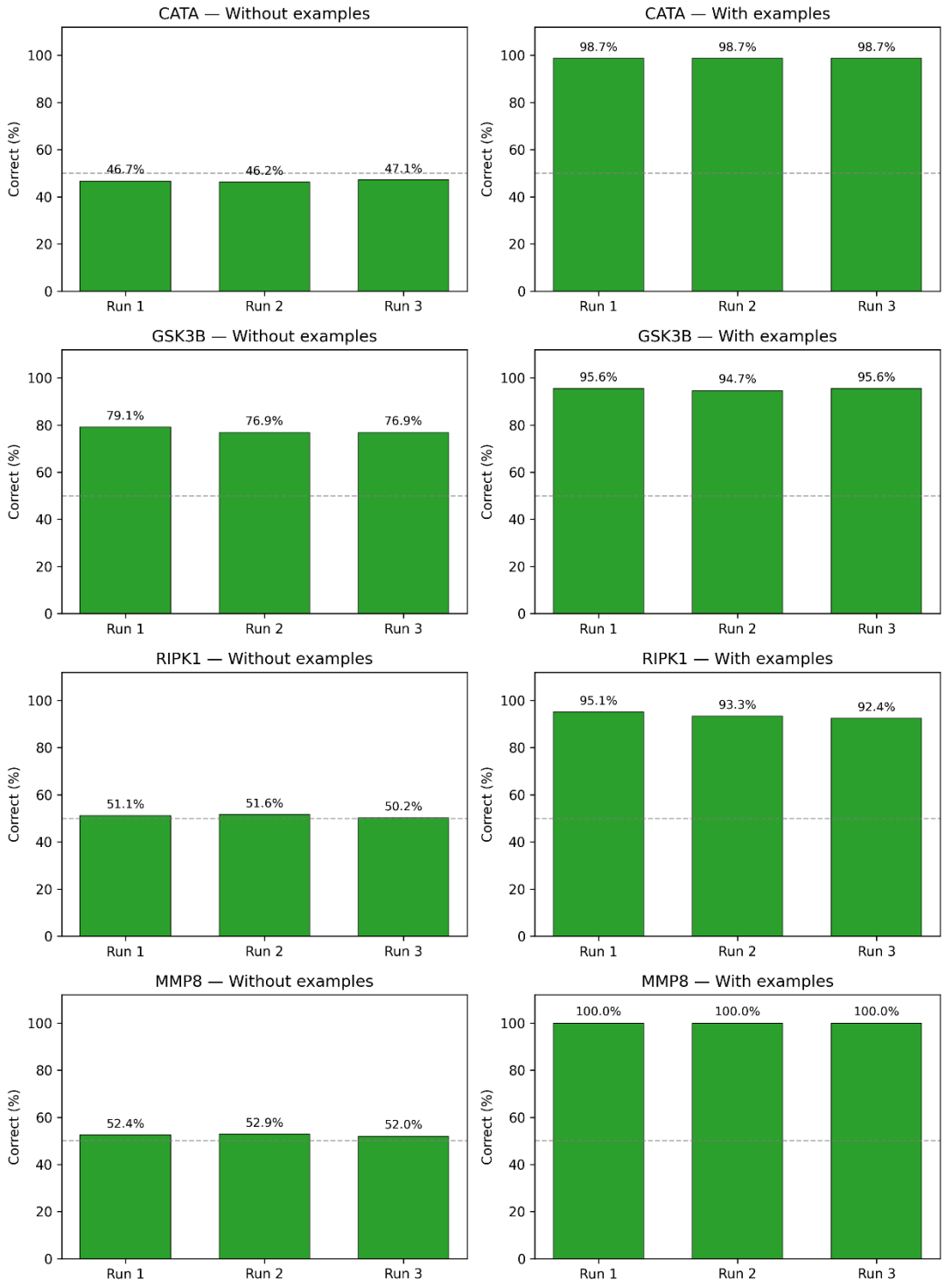


Fig. S3: Performance of correctly classifying the more active compound of 4 datasets with active-inactive pairs (15 molecules in each class) with three repetitions. Left plots show the results for ICL prediction without any additional information, right plots show the results for ICL-prediction with context data.

Iterative design strategy increases compound potency, overcoming predictive limitations

The iterative optimization approach implemented in this study led to a clear improvement in the experimental activity of the designed compounds, as evidenced by the shift in their pIC_{50} values in Fig. S13, with 9 out of the 10 compounds exhibiting higher activities compared to their original counterparts. This confirms the practical effectiveness of our method in identifying more potent molecular candidates.

While computational predictions, especially those generated by the LLM, tended to overestimate the experimental activity, the experimental results of the optimized compounds remained promising. Importantly, 60% of the experimentally validated optimized compounds satisfied the predefined activity threshold, underscoring that the approach successfully identified genuinely potent molecules within the desired range despite predictive inaccuracies. From Fig. S4 it is also observed that the ensemble of descriptor-specific oracles provided predictions that were somewhat closer to experimental observations compared to the LLM, yet the moderate correlation between these computational approaches highlights their complementary nature.

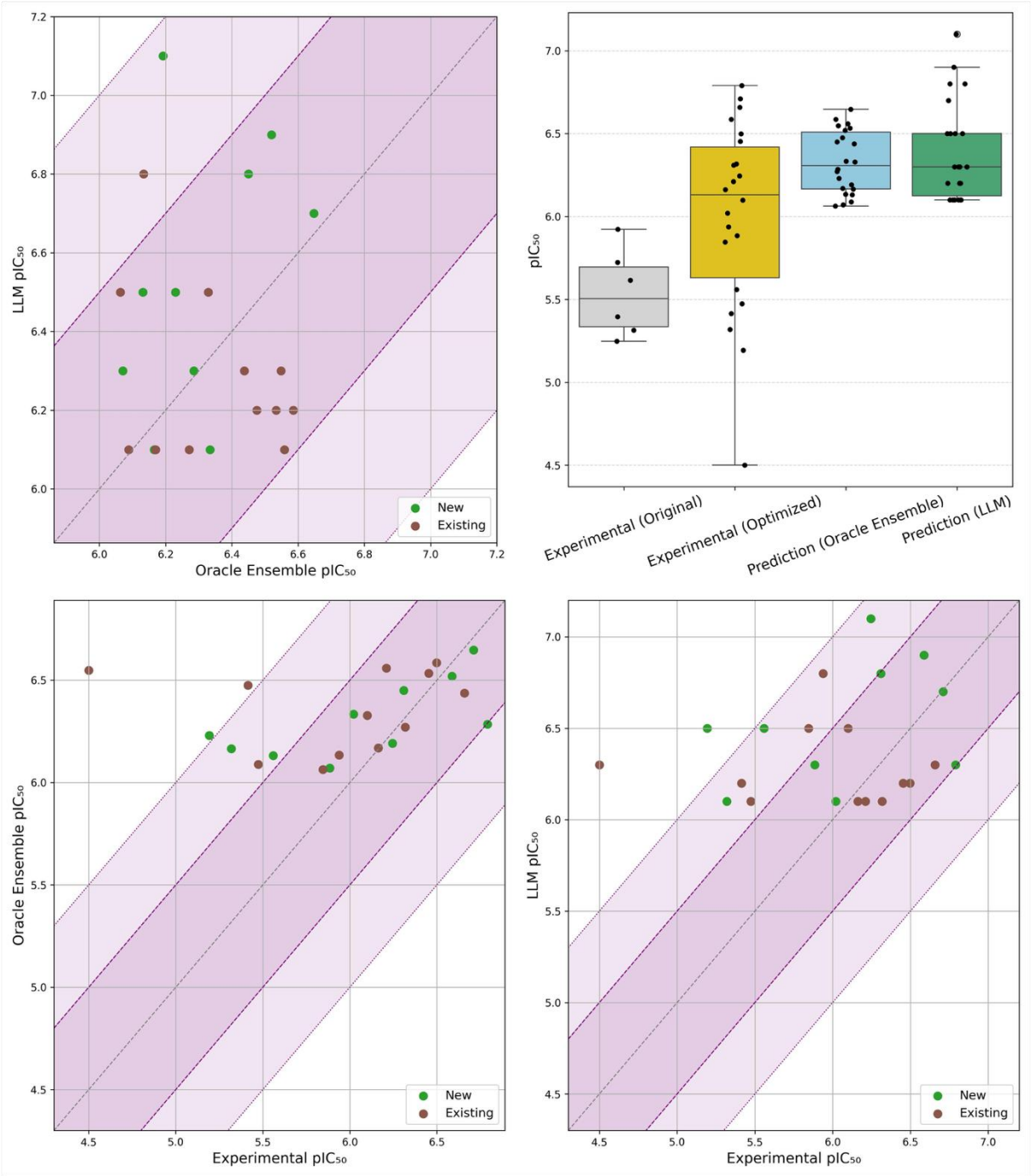


Fig. S4: Predictive alignment and experimental outcomes of the iterative design process. Upper left: Scatter plot showing the LLM oracle ensemble-predicted pIC_{50} values for the selected compounds. Upper right: Boxplot comparison of activity distributions for the original set of compounds, the selected, and predictions from both the LLM and oracle ensemble. Gains in experimental activity is achieved despite the limited agreement between approaches. Lower left: Scatter plot showing oracle ensemble-predicted pIC_{50} versus experimental pIC_{50} values for synthesized compounds. Lower right: Scatter plot showing LLM predicted pIC_{50} versus experimental pIC_{50} values for synthesized compounds.

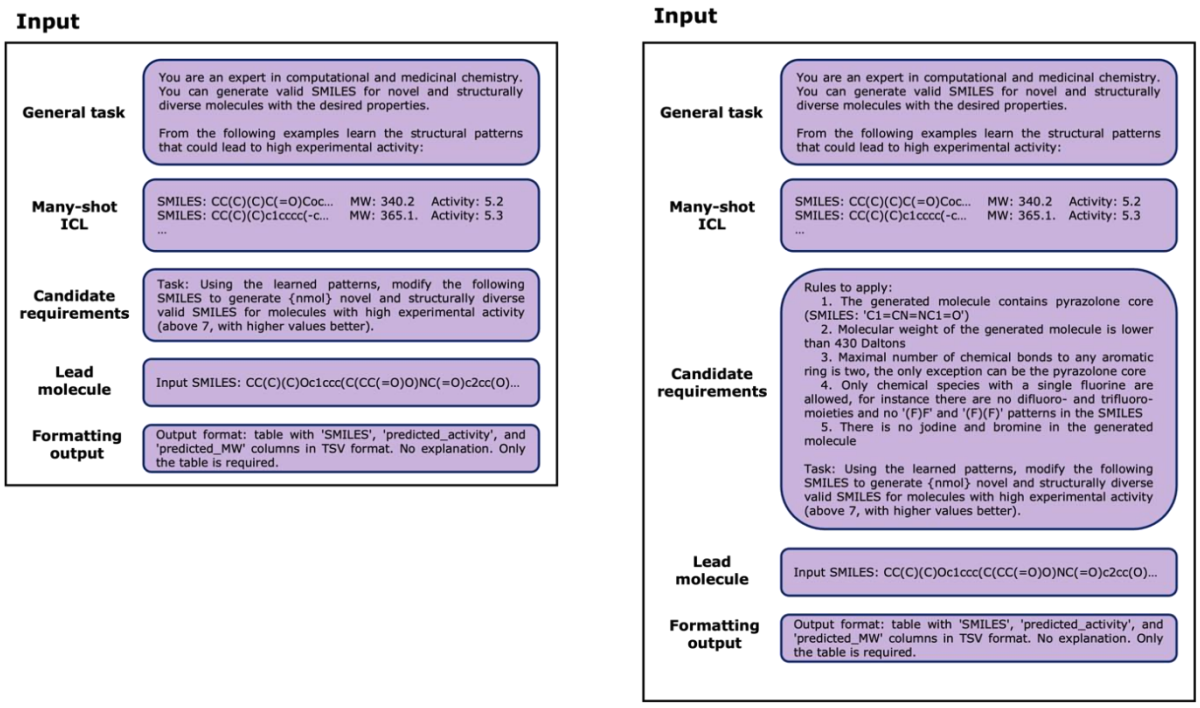


Fig. S5: Non-rule and rule-based prompt templates used for the generation of new candidates through ICL. By providing clear instructions and a larger and diverse pool of examples, the LLM learns underlying structural patterns that lead to the molecular property value. We need to explicitly indicate every detail around the task that is to be conducted. This includes the property requirements, the starting molecule, and how the output should be formatted. By learning on the examples provided in the prompt, the LLM generates new unseen candidates which should match the established criteria.

Reproducibility

We conducted 200 independent inferences using both rule-based and non-rule-based prompts and analyzed the frequency of individual generated SMILES retrieved for each query molecule. The default temperature value of 1 was maintained. The temperature parameter controls the randomness of the output: lower values make the model more deterministic, while higher values make it more creative and diverse. The comparison of both prompts is shown in Table S1, demonstrating that a significant majority of retrieved SMILES are valid molecular representations.

The reproducibility of the LLM design was evaluated by the fraction of unique molecules proposed through multiple queries of the LLM. Table 1 shows a partial overlap between individual evaluations of the same prompt, with approximately 30% of all designed molecules being unique for the non-rule-based prompt and 5-15% for the rule-based variant. Hence, in both cases, the majority of molecules were proposed more than once in our 200 test runs.

Additionally, we explored how many molecules from both prompting strategies would fulfill the criteria given as constraint in the rule-based prompt. Hence, all molecules from both prompts were checked if all rules from the rule-based prompt were met. Molecules generated by the non-rule-based prompt consistently missed to meet one or more criteria for designed molecules. Conversely, the rule-based

prompt achieved high reliability in meeting design constraints for starting from most query compounds (see Table 1). Lastly, the molecular pools generated by both prompts formed completely disjoint sets (see column overlap in Table 1).

Effect of LLM's temperature parameter on sampling

Since the default temperature parameter for inference in the Claude Sonnet model is 1, we tested the performance of this LLM with lower values, ranging from 0.05 to 0.8. The statistics are presented in Table S1. We found that the temperature parameter value does not significantly affect the fraction of valid and rule-compliant outputs. However, the diversity, measured as the number of unique SMILES, increases monotonically with rising temperature. This observation aligns with the interpretation of the temperature parameter as a control for the randomness and creativity of the LLM outputs (Table S1). Impact of LLM Temperature Parameter on generated SMILES. Statistics are based on a fixed prompt tailored to the single query compound, "query 1", executed 50 times for each temperature.

Table S1: Diversity and validity when applying different temperature factors ranging from 0 to 1.

	non-rule-based prompt			rule-based prompt		
Temperature	valid [%]	Unique [%]	rule compliant [%]	valid [%]	Unique [%]	rule compliant [%]
0.05	100	22.8	0	100	12	59
0.2	100	25	0	99.4	13.6	57.8
0.4	100	32.6	0	99.6	15	58.6
0.6	100	40.2	0	99.6	15.4	59
0.8	99.2	46.6	0	97.8	19.6	62.2
1	99.8	45.8	0	99	22.4	60.2

Prevalence of Validated Designed Molecules in Pools Generated by Many-Shot ICL

We analyzed the experimentally validated molecules generated by the iterative Many-Shot ICL design to see how often they appeared in compound sets created using the simpler, one-step Many-Shot ICL method. Table S2 shows the frequency of the synthesized molecules ID 11-20 across 200 independent runs for each query molecule.

Interestingly, none of the designed molecules appeared in compound sets generated using non-rule-based prompts, suggesting that these prompts explored a different region of chemical space. In contrast, with rule-based prompts, the frequency of synthesized molecules ID 11-20 varied greatly, and some molecules appeared relatively often.

These results indicate that the iterative Many-Shot ICL approach is more effective at exploring broader chemical space, thereby increasing the chance of identifying promising hit compounds.

Table S2: Frequency of the selected compounds in the 200 replicated LLM test queries using the same prompt.

id	Non-rule-based prompt	Rule-based prompt
ID11	0	7
ID12	0	25
ID13	0	80
ID14	0	5
ID15	0	0
ID16	0	95
ID17	0	140
ID18	0	0
ID19	0	9
ID20	0	8

LLMs produce new chemical moieties

We analyzed the chemical moieties present in the query compounds and the SAR-data and compared their occurrence in the compounds generated by the LLM. The graphical output of this analysis is presented in Fig. S6. One of the key findings is that the analogs designed using the non-rule-based prompt completely lack the pyrazole moiety, which is a defining feature of the chemical series from which the starting point compounds are derived.

Furthermore, the analysis indicates that the designed compounds predominantly incorporate moieties found in the SAR-data, with some exceptions observed in both the rule-based and non-rule-based prompt outputs. For instance, the thiazole heterocycle, commonly present in the example compounds, was not found in the compounds generated by the rule-based prompt, although it was frequently utilized in those derived from the non-rule-based prompt.

Notably, the LLM also introduced novel moieties that were absent from both the query compounds and SAR-data. These new fragments (including the thiophene ring, cyclohexane ring, piperidine ring, cyclobutyl group, and triazine ring) were detected at lower frequencies but demonstrate the model's capacity to enrich the pool of generated compounds with structurally diverse and potentially novel motifs.

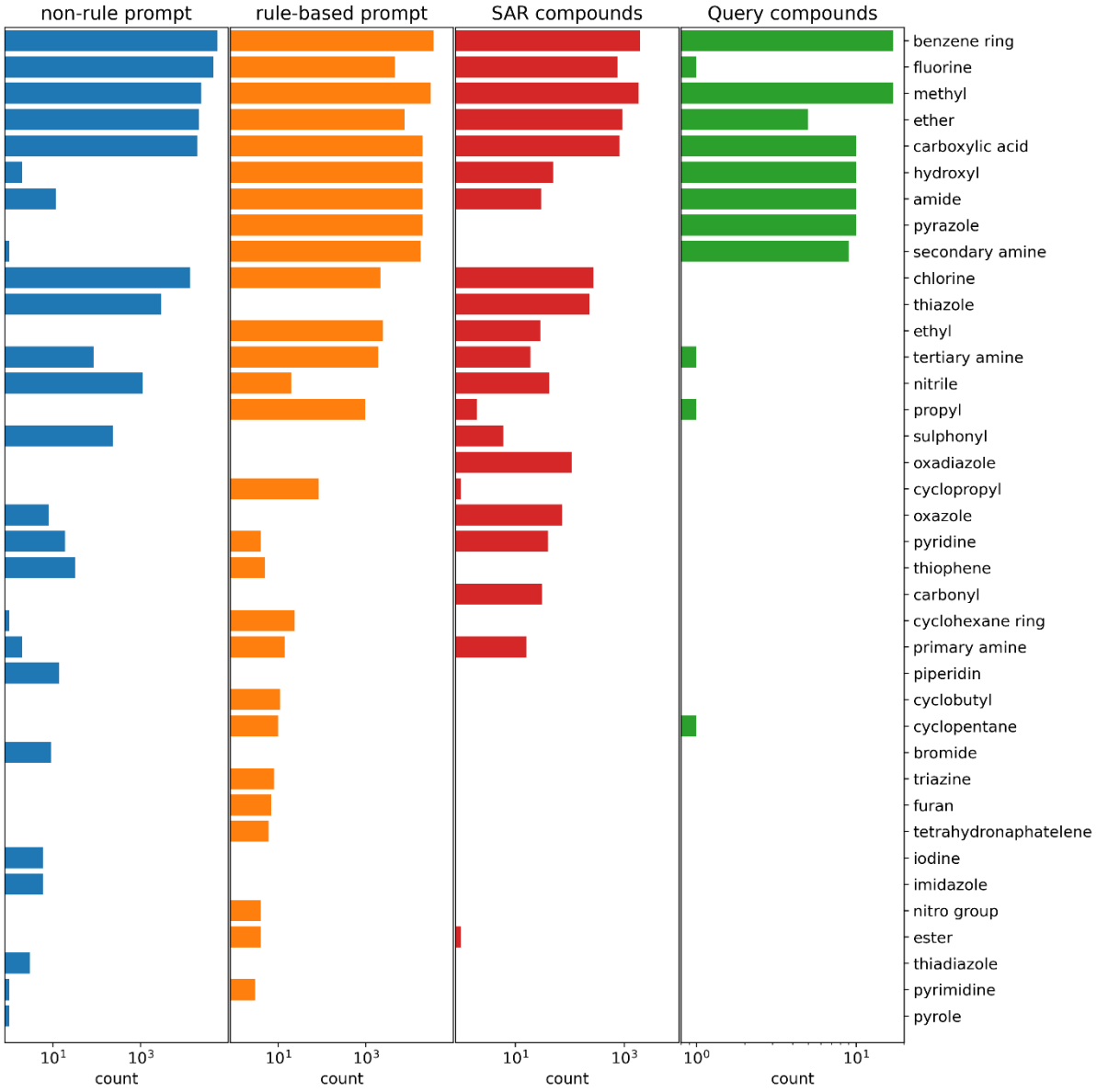


Fig. S6: Analysis of chemical moieties in molecules generated using both rule-based and non-rule-based prompts, alongside the example compounds and query molecules provided to guide the LLM.

The incorporation of specific chemical moieties in the designed compounds can be effectively modulated using rule-based prompts

To evaluate the impact of additional constraints, we introduced rules into the prompt that explicitly prohibited the inclusion of fluorine and carboxylate groups in the LLM-generated compounds. As shown in Fig. S7, the application of these rules successfully eliminated the presence of the targeted moieties; neither fluorine nor carboxylate groups were detected in any of the generated compounds. Interestingly, the frequencies of other chemical groups also appeared to be slightly affected. This may be attributed to their co-occurrence with the prohibited moieties in the SAR-data, or alternatively, to a shift in the model's intrinsic bias induced by the presence of additional constraints during the molecule generation process.

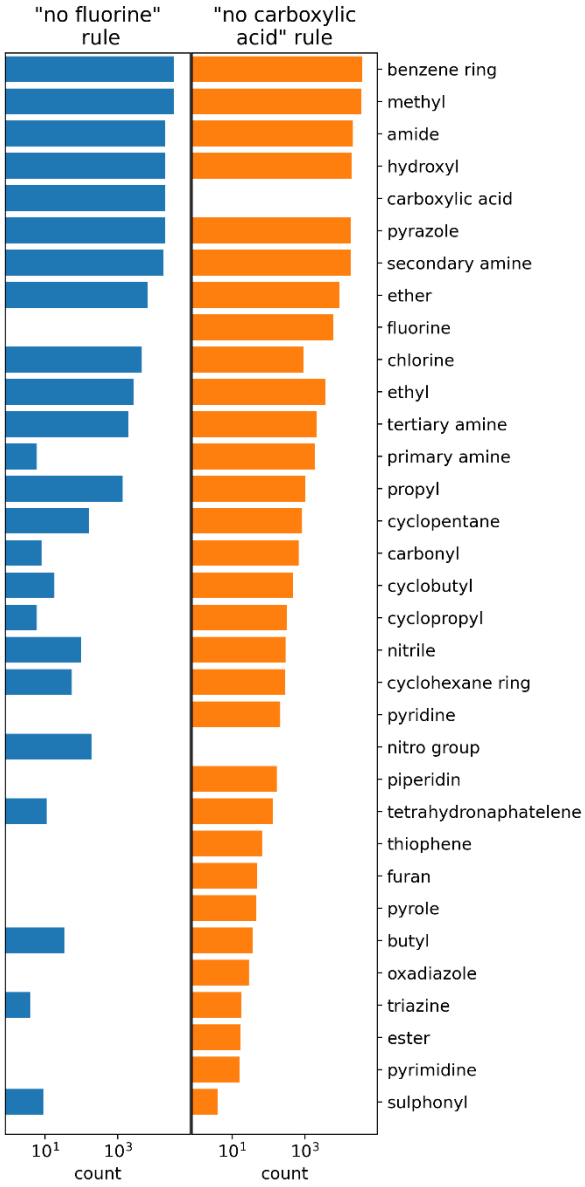


Fig. S7: Analysis of chemical moieties in molecules generated using additional rules for avoiding fluorine and carboxylic groups in the LLM-designed analogs.

Distribution of metrics for compound accessibility

For comparing and illustrating the generated ensembles before and after filtering and selections steps, we calculated the distribution of the QED-score (as a metric of drug likeliness) as well as the SAScore (as a metric of compound synthesizability) (see Fig. S8).

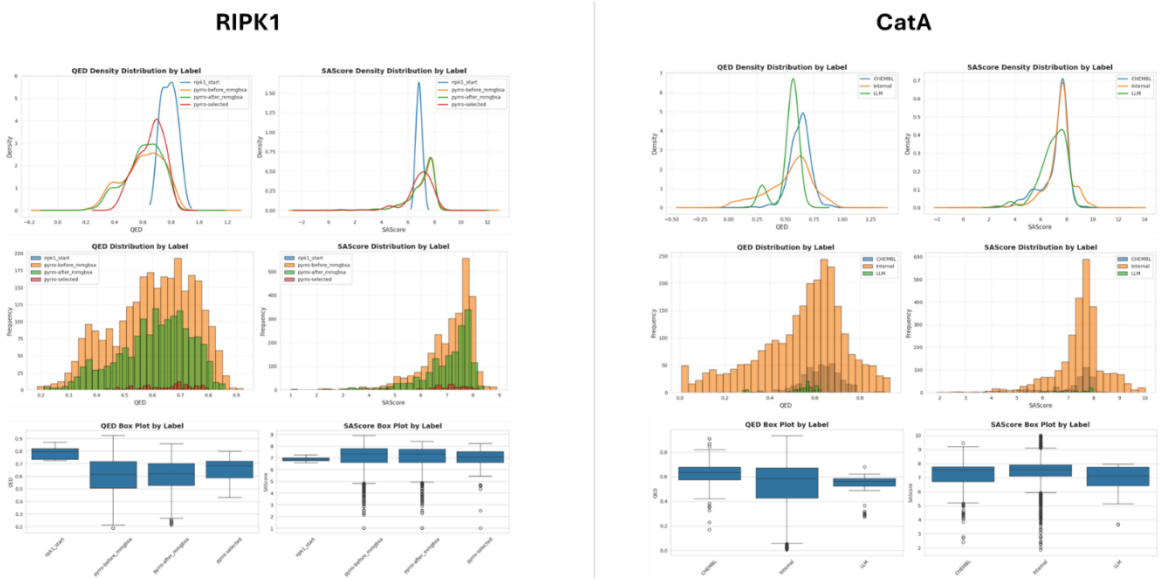


Fig. S8: Distribution of QED and SAScore for starting molecules and generated molecules. Left panel: RIPK1; starting molecules (blue) are compared to a) generated molecules after unwanted fragment filtering (orange), b) after docking and MM-GBSA selection (green), and c) the final selection of compounds for synthesis (red). Right panel: CatA; internal query molecules (orange) are compared to the LLM design molecules (green) as well as CatA compounds already available in the ChEMBL-database.

Novelty of generated molecules and similarity to query compounds

To demonstrate the novelty of the generated compounds, we calculated the Tanimoto similarity with respect to the ChEMBL36 database with a Morgan type fingerprint of radius 2 and a bit size of 2048. Structural verification confirms that none of the compounds possess an exact match in the public database. Fig. S9 depicts the distribution of maximum similarity to any of these compounds, revealing that 94 out of the 95 generated molecules for CatA exhibit a value below 0.7 and being the mean of 0.538 and an even lower similarity for 119 generated molecules for RIPK1. This significant structural dissimilarity confirms that our pipeline can successfully generate novel compounds rather than rediscovering existing bioactive scaffolds.

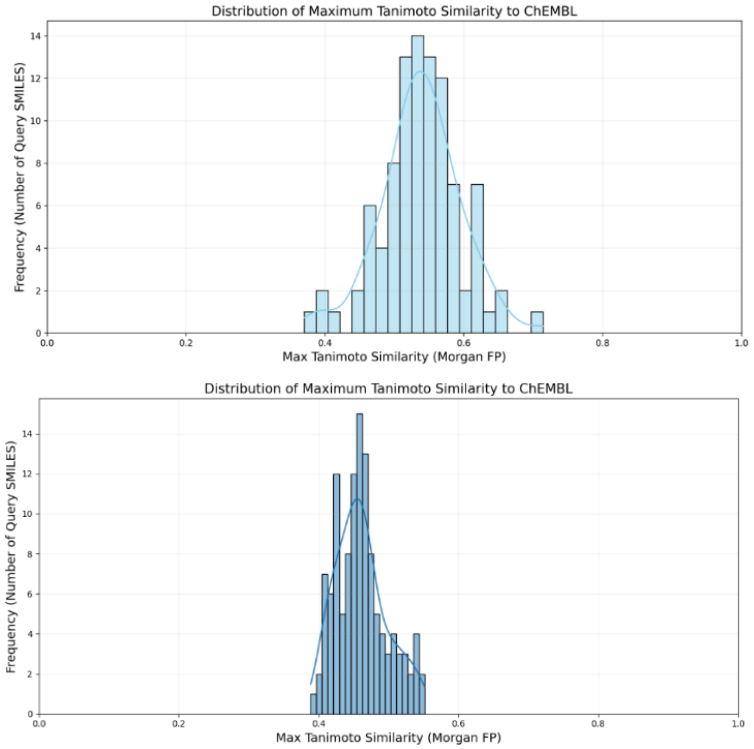


Fig. S9 Distribution of Tanimoto similarity of generated molecules with respect to ChEMBL36. Upper panel: 95 generated molecules for CatA, lower panel: 119 selected generated molecules for RIPK1.

In addition, we calculated the Tanimoto similarity of the synthesized molecules towards the starting points (see SI Fig. S10 and S11). The generated library maintains a controlled similarity profile relative to the starting points (typically ranging between 0.4 and 0.7), where core pharmacophoric features are preserved while peripheral decorations are explored. Notably, the experimentally validated hits, which represent a subset of the 95 generated compounds, showed a mean similarity to their respective queries of 0.594, demonstrating that the molecules are distant enough from the queries to be considered distinct analogs, yet close enough to retain the targeted binding modality.

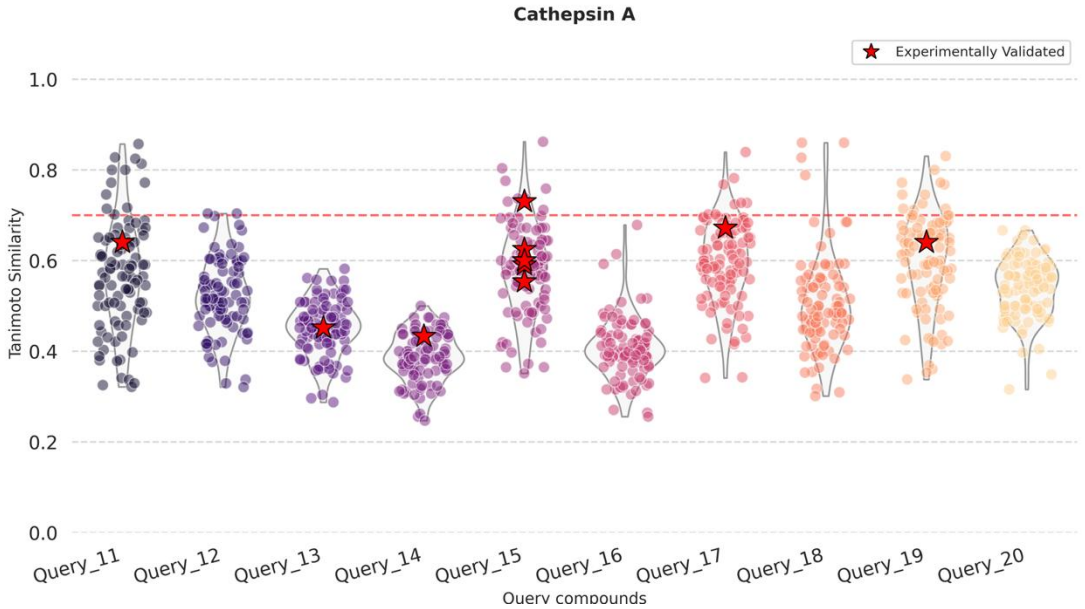


Fig. S10 Tanimoto similarity distribution for generated molecules for CatA with respect to each query. The synthesized compounds are marked by a red star.

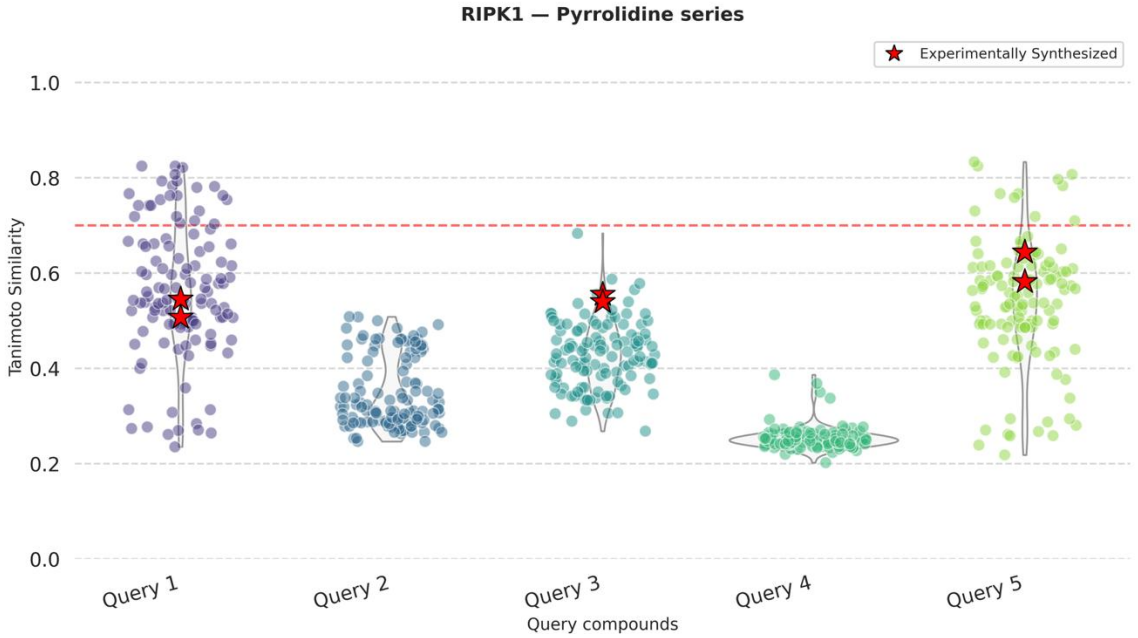


Fig. S11: Tanimoto similarity distribution for generated molecules for RIPK1 pyrrolidine series with respect to each query. The synthesized compounds are marked by a red star.

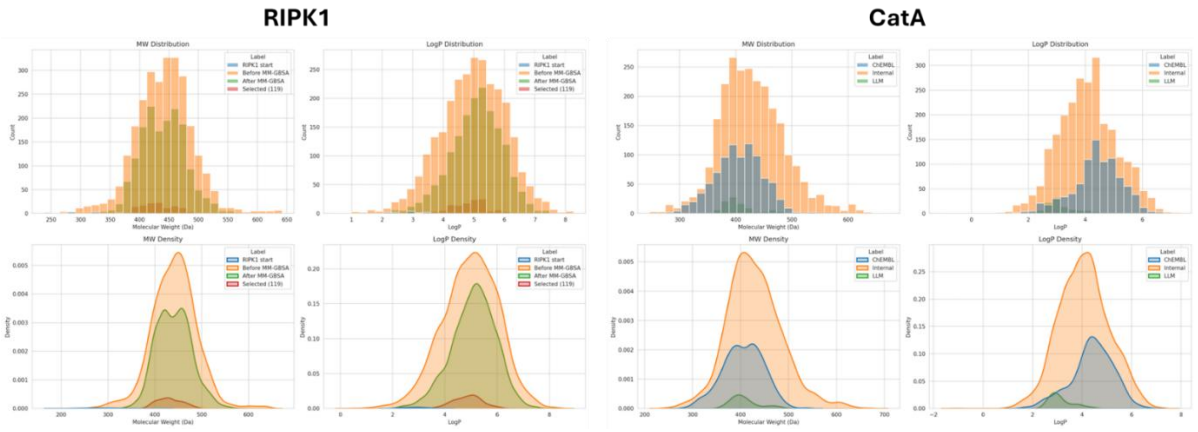


Fig. S12: Distribution of physico chemical properties molecular weight and logP (calculated by rdkit) for Left panel: RIPK1; starting molecules (blue) are compared to a) generated molecules after unwanted fragment filtering (orange), b) after docking and MM-GBSA selection (green), and c) the final selection of compounds for synthesis (red). Right panel: CatA; internal query molecules (orange) are compared to the LLM design molecules (green) as well as CatA compounds already available in the ChEMBL-database.

LLM proposals show strong enhancement of activity across new synthesized and existing compounds

In the CatA project, out of a total of 95 LLM-based candidates, twelve of the designed compounds were found in internal historical data, despite the LLM never having seen them, and eleven showed improved activity over their queries. Of the ten novel compounds that were synthesized and tested, nine out of ten exhibited higher pIC_{50} values. Altogether, 20 of the 22 designs (91%) demonstrated measurable potency gains, confirming that the LLM consistently proposes substitutions and structural modifications that translate into activity improvements.

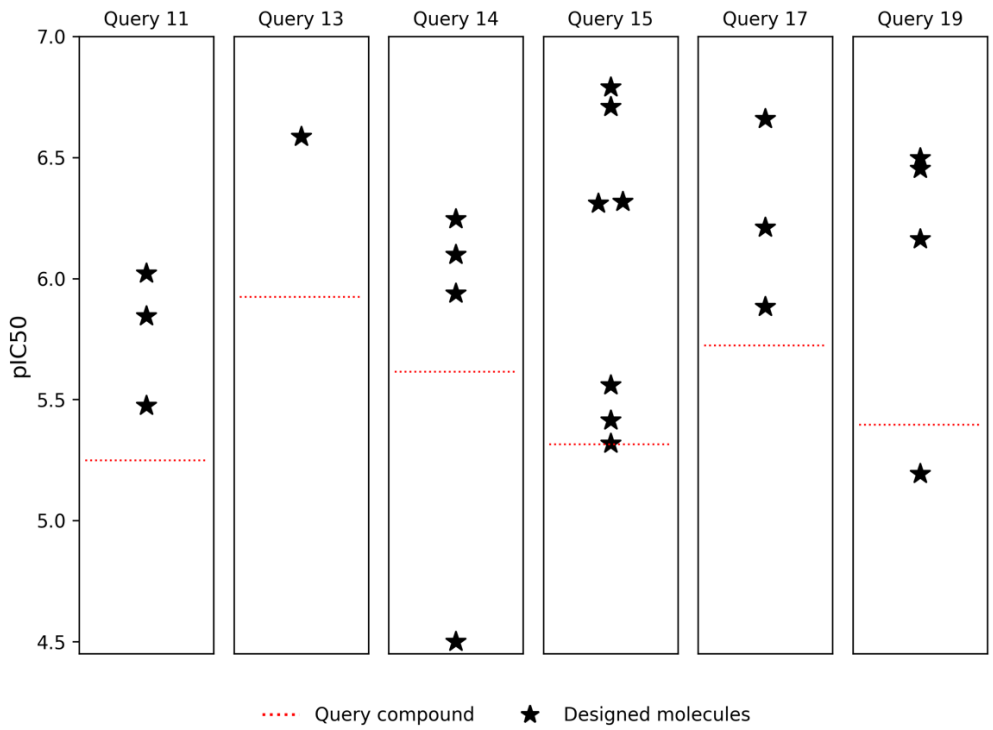


Fig. S13: Experimental activities of all CatA LLM-designed compounds compared to starting queries as a baseline.

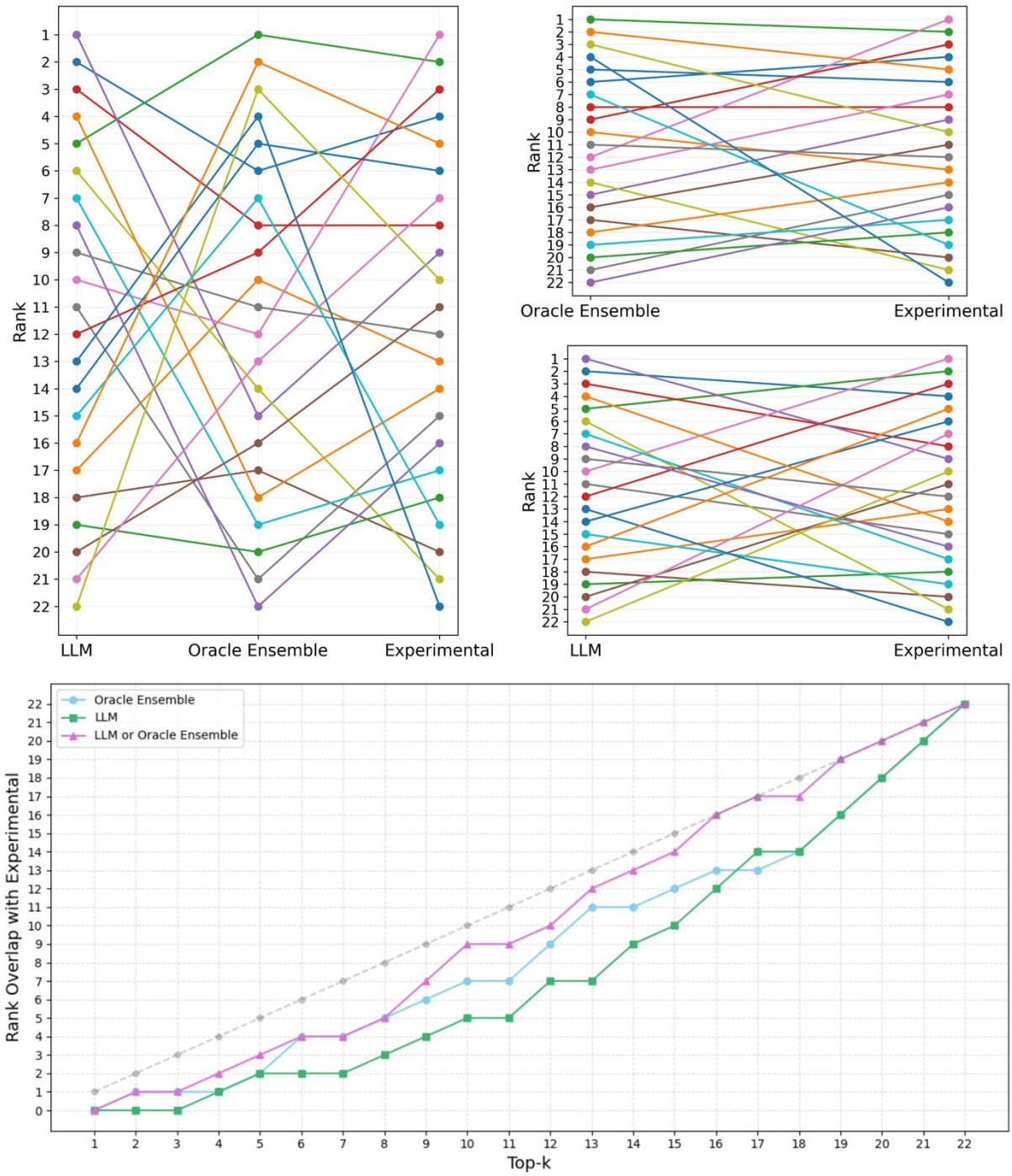
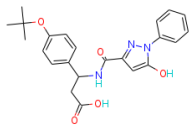
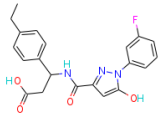
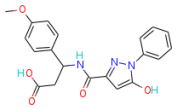
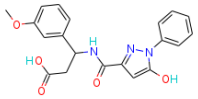
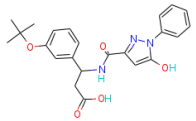
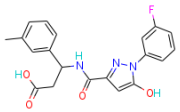
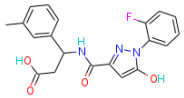
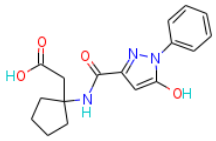
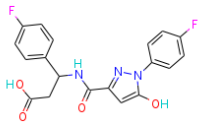


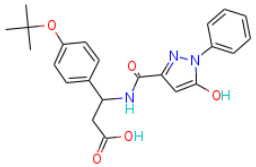
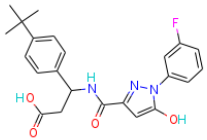
Fig. S14: Compound Ranking based on predicted and experimental activities. Each colored line represents one candidate molecule, highlighting how the position changes from in silico to real experimental ranking.

Chemical reasoning for molecules designed by the LLM

Table S3: Some examples from designed compounds for the CatA example with the proposed reasoning of the LLM.

Example	Starting point	Designed molecule	Predicted activity	Experimental activity	Reasoning provided by the LLM	Conclusion
1			6.2	6.5	The tert-butoxy group is replaced with an ethyl group: maintaining lipophilicity. The addition of a fluorine atom on the phenyl ring of the pyrazole could improve binding through electrostatic interactions. This modification is supported by the presence of fluorinated compounds with high activity in the example set.	Systematic exploration of t-Butoxy group is suggested. The position is moved around the ring to explore different regio-isomers.
			6.0	6.5	The tert-butoxy group is replaced with a methoxy group: which may improve solubility while maintaining some lipophilicity. The pyrazole structure is maintained. The slight decrease in predicted activity accounts for the smaller substituent: but the modification is still likely to be beneficial overall.	
			6.0	6.2	The optimization replaces the para tert-butoxy group with a meta methoxy group. This change reduces lipophilicity and steric bulk while maintaining hydrogen bond accepting capability. The meta position may allow for better fit in the binding pocket. The predicted activity of 6.1 seems reasonable: but I've adjusted it to 6.0 to account for the significant structural change.	

2		 ID 11	6.2	6.8	The tert-butoxy group is replaced with a methyl group: reducing size but maintaining some hydrophobicity. The addition of fluorine to the phenyl ring on the pyrazole and the rearrangement of substituents may create new binding interactions and improve metabolic stability.	Incorporation of beneficial F-atom which can also be seen in reference compounds from the SAR-data
		 ID 12	6.5	6.7	The tert-butoxy group was replaced with a methyl group: and a fluorine was added to the phenyl ring. These changes likely improve metabolic stability while maintaining lipophilicity. The core structure is preserved. The fluorine addition may enhance binding affinity: as seen in similar examples.	
3		 ID 15	7.3	6.2	The cyclopentyl ring was replaced with a para-fluorophenyl: likely improving π-π-stacking interactions. The addition of fluorine to both aromatic rings is consistent with the trend of increased activity in fluorinated compounds seen in the example set. The more linear structure may better fit the binding pocket. The prediction is increased based on these favorable modifications.	The LLM correctly suggests to replace the aliphatic acid by an aromatic acid.

4		 ID 20	6.5	5.2	The optimization replaces the tert-butoxy group with a tert-butyl group and introduces a meta-fluorine on the phenyl ring. These changes may improve metabolic stability and binding affinity. The predicted activity increase from 5.4 to 6.5 is significant but plausible: as similar modifications in the example molecules often lead to improved activity.	The suggestion from the LLM was wrong. The reasoning can help chemists to judge the suggested change.

Candidate pool behavior during iterative process

The molecular optimization workflow implemented in this study is based on an iterative, self-guided strategy consisting of 10 rounds. In each iteration, a fixed number of designed molecules are independently generated for every initial query compound. In the very first iteration, we specifically included biaryl in the prompt presented to the LLM. This strategic decision leverages the chemical knowledge embedded in these biaryl, providing useful structural cues to guide initial candidate generation. However, this chemical series was intentionally excluded from the evolving candidate pool to ensure the iterative process remained focused on refining and optimizing compounds structurally similar to the designated query compounds.

Throughout each iteration, newly generated molecules were integrated into the evolving candidate pool only if their LLM-predicted activities surpassed a dynamically adjusted threshold, defined as the mean predicted activity of the preceding iteration. To rigorously control pool quality, systematic sanity checks were performed. Each newly generated SMILES was canonicalized, eliminating potential representation inconsistencies. Duplicates arising from within or across previous iterations, as well as chemically invalid or problematic SMILES, were identified and discarded. These quality-control procedures ensured structural validity, chemical feasibility, and uniqueness within the pool.

As the iterative optimization proceeded, both the pool size and mean predicted activity progressively increased, reflecting the selective retention of highly promising candidates. However, a natural convergence limit was observed in later iterations, driven by increasingly strict activity thresholds. As fewer molecules satisfied these stringent criteria, the rate of inclusion slowed, and the pool stabilized in size and activity. This observed convergence in Fig. S15 can be attributed primarily to saturation of the chemical space around the initial scaffolds, the predictive ceiling of the LLM model, and practical constraints on achievable potency improvements.

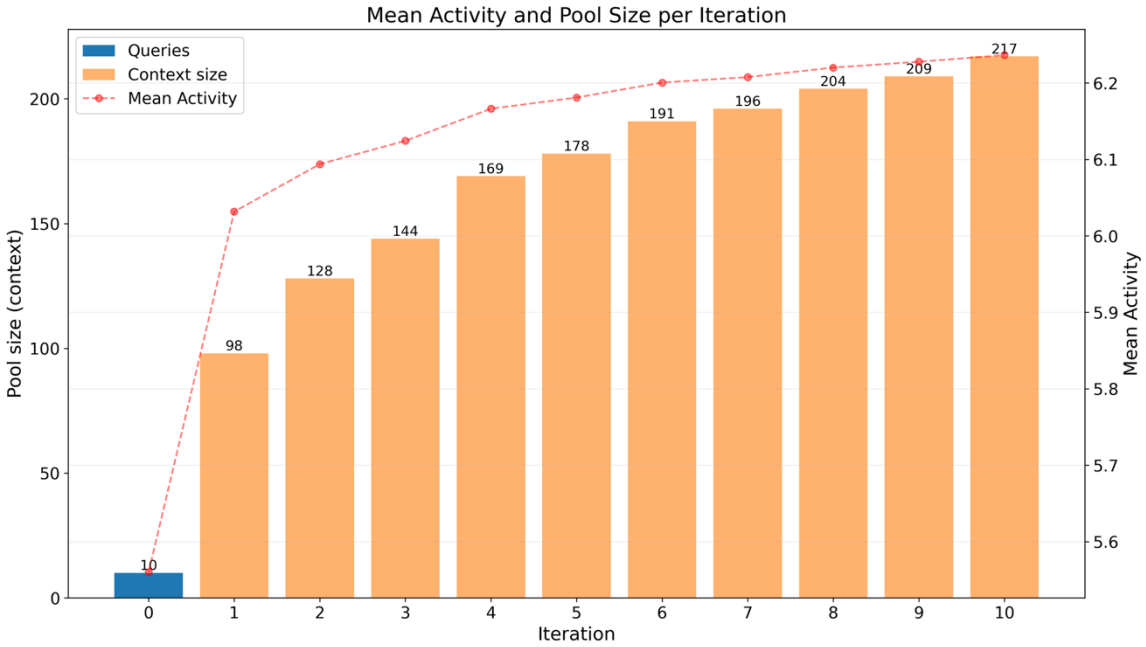


Fig. S15: Iterative expansion and enrichment of the candidate pool. An increase is observed in both the pool size and its mean activity. With each round, new candidates generated by the LLM that exceed the activity threshold are incorporated, leading to a steady growth. The trend reveals an initial rapid expansion and enrichment phase, which is followed by a clear convergence, as stricter activity threshold limits the addition of new molecules.

Ultimately, the iterative self-guided strategy successfully balanced broad chemical exploration with targeted structural refinement. The incorporation of biaryl solely within the initial prompt, followed by their intentional omission from the evolving context, effectively guided the optimization toward structurally relevant chemotypes. These decisions, combined with filtering and dynamic thresholding, yielded a refined candidate set enriched in structurally relevant and potential high activity compounds.

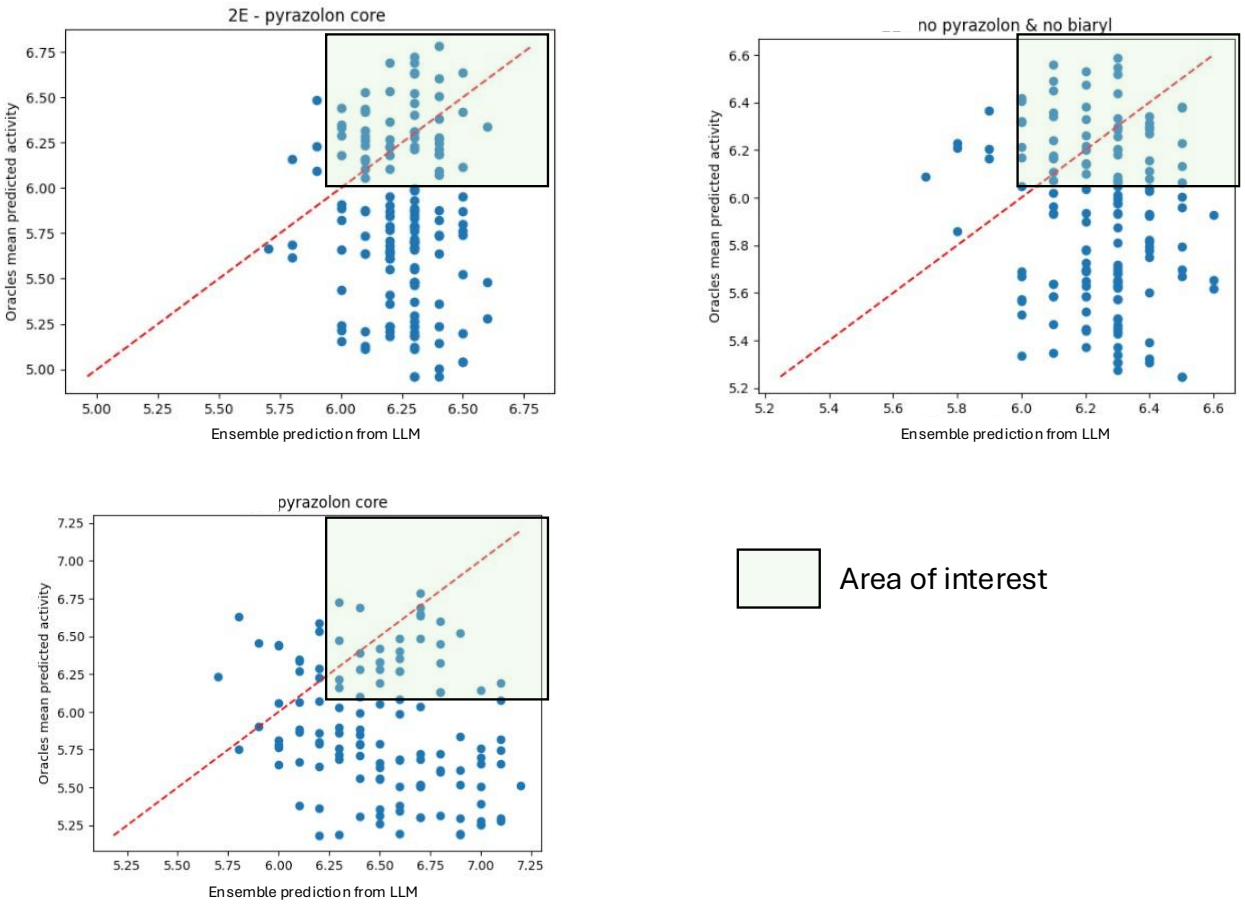


Fig. S16: Comparisons of predictions from ensemble of ML models and the LLM predictions. The upper right corner was used as selection of designed molecules.

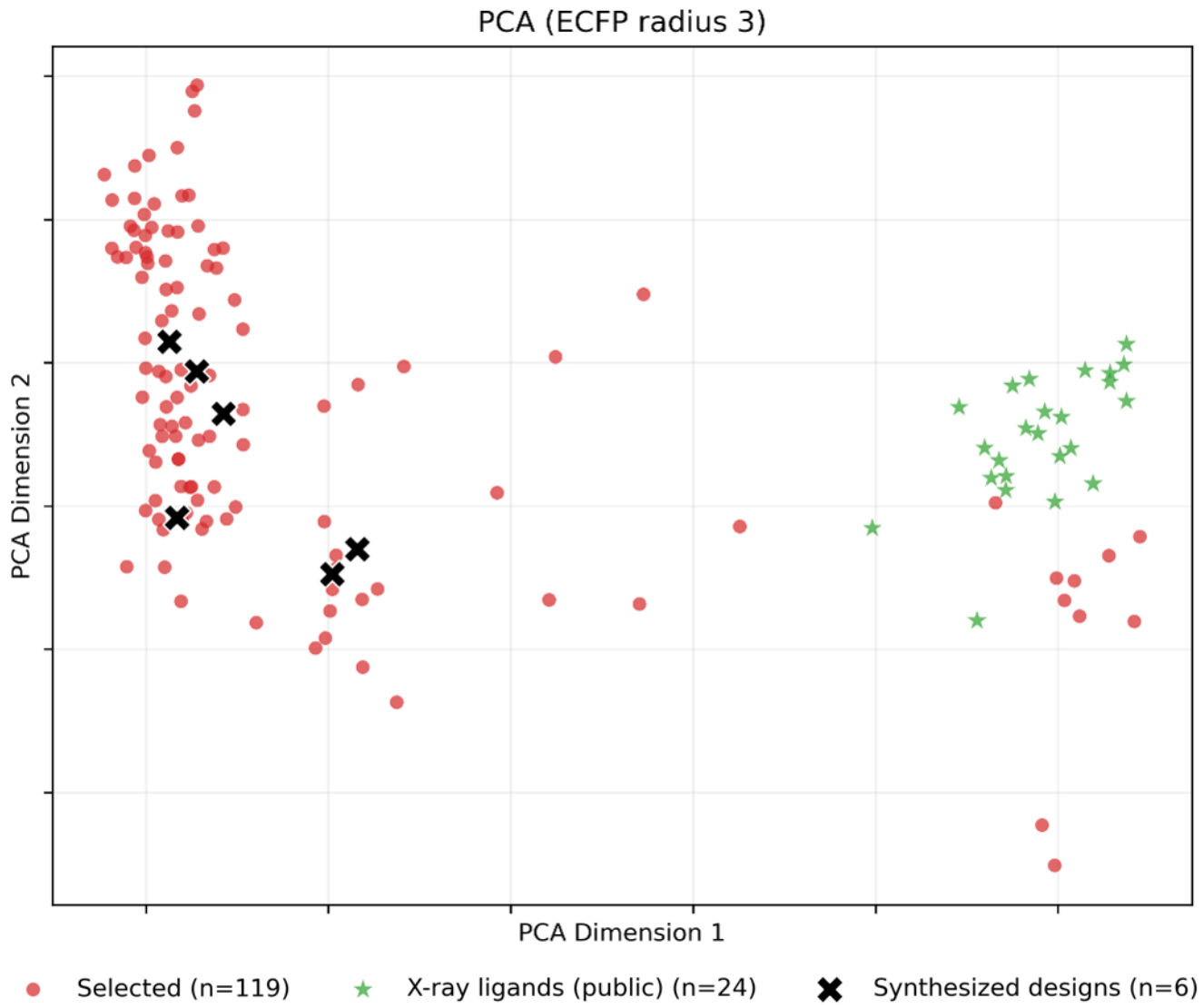


Fig. S17: Principal component analysis of 119 LLM-generated pyrrolidine derivatives using ECFP circular fingerprints (radius 3, 2048 bits) with 24 known reference RIPK1 type III binders and 6 synthesized compounds from LLM design.

Descriptors

For the predictive activity models, we trained models using three different types of features, namely circular fingerprints, RDKit descriptors, and Mol2Vec molecular features.

Circular Fingerprints are generated using a hash function that transforms the local environment *E* of each atom into a bit in a fixed-size vector. For each atom, the environment is defined by the type of atom and its neighbors up to a specified radius r:

$$X_{cf}[i] = \bigoplus_{\text{atom } j \in \text{Atoms}(s)} \text{Hash}(\text{E}(j,\text{r}))$$

Here, Xcf [i] is the i-th bit of the fingerprint vector, L denotes a bitwise OR operation across all atoms, and E(j, r) represents the substructure around atom j within the given radius. Hash is a function that maps each unique substructure to a bit position.

RDKit computes a set of predefined descriptors for each molecule. These descriptors are functions of the molecular structure, such as molecular weight, logP, and the count of various types of bonds and atom types. The generation of a descriptor vector can be mathematically represented as:

$$X_{rd} = [f_1(s),\, f_2(s), \ldots, f_n(s)]$$

Where $f_k(s)$ is the k-th descriptor function applied to the SMILES string s, and n is the total number of descriptors calculated by RDKit.

Mol2Vec converts each molecule into a vector by first tokenizing the SMILES string into meaningful chemical fragments and then mapping each fragment to a vector in a pre-trained embedding space:

$$X_{mv} = \frac{1}{m}\sum_{k=1}^{m} \text{Vec}(\text{Token}_k)$$

In this equation, m is the number of tokens in the SMILES string s, Token_k is the k-th token, and $\text{Vec}(\text{Token}_k)$ is the vector representation of Token_k from a pre-trained embedding model. The final feature vector X_{mv} is the average of all token vectors, providing a dense representation of the molecule.

Each feature set X_{cf} , X_{rd}, and X_{mv} is then used as input to an ensemble of ML models, which are trained to predict molecular activity a’ based on these distinct representations:

$$M(X) \rightarrow a'$$

Compound synthesis

RIPK1

Analytical methods

NMR Method: spectra were recorded on a Bruker AVANCE NEO 700 spectrometer operating at a proton frequency of 700.10 MHz and a ^{13}C-carbon frequency of 176.0 MHz. The instrument was equipped with a 5 mm TCI cryoprobe. Or a Bruker AVANCE II 400 spectrometer operating at a proton frequency of 400.32 MHz and a ^{19}F frequency of 376.5 MHz and equipped with a 5mm PABBI 1H/1D-BB probe with z-gradient.

LCMS Method 1: Retention time and mass detection were done on a Waters Acquity UHPLC system coupled with a Waters SQD mass detector. The injection volume was 1.0 µl. Molecular weights are given in gram per mol [g/mol], detected masses in mass per charge [m/z]. Gradient: 98% H_2O (0.05% formic acid)/2% acetonitrile (0.035% formic acid) for 0.2 min, then from 98% H_2O (0.05% formic acid) to 98% acetonitrile (0.035% formic acid) in 3.6 min, then 98% acetonitrile (0.035% formic acid) for 0.5 min, flow rate: 1.0 ml/min, column: 2.1×50 mm Waters ACQUITY UPLC BEH C18, 1.7 µm, 55°C. UV data: retention time ad λ = 220 nm given in min; MS data: ES+ ionization, m/z given as [M+H]+ unless otherwise noted.

Intermediate ((S)-2-(3,5-difluorophenyl)pyrrolidin-1-yl)((S)-3-hydroxypyrrolidin-1-yl)methanone (IIIa) and ((S)-2-(3,4-difluorophenyl)pyrrolidin-1-yl)((S)-3-hydroxypyrrolidin-1-yl)methanone (IIIb):

Synthesis of Intermediate IIa

Step 1a

To a solution of (2S)-2-(3,5-difluorophenyl)pyrrolidine (300 mg, 1.64 mmol) in CH_2Cl_2, CDI (292.1 mg, 1.1 eq) was added, and the reaction was stirred at rt for 1h. Then the mixture was washed with water and the organic phase dried over Na_2SO_4, filtered and evaporated to yield (2S)-2-[(3,5-difluorophenyl)pyrrolidin-1-yl]-imidazol-1-yl-methanone (439 mg, 97%) which was used in the following step without further purification.

^{1}H NMR (400.23 MHz, DMSO-d6) δ ppm 8.20 (br s, 1 H), 7.62 (br s, 1 H), 7.09 (m, 3 H), 7.00 (br s, 1 H), 5.09 (br t, *J*=7.34, 7.34 Hz, 1 H), 4.03 (m, 1 H), 3.68 (m, 1 H), 2.42 (m, 1 H), 1.92 (m, 2 H), 1.74 (m, 1 H). LCMS (Method1): RT 1.36 min, 278 [M+H]+.

Step 2a

To a solution of (2S)-2-[(3,5-difluorophenyl)pyrrolidin-1-yl]-imidazol-1-yl-methanone (439 mg, 1.58 mmol) in acetonitrile (3.2 mL), iodomethane was added (0.15 mL, 1.5 eq) and the mixture was stirred at RT overnight. LCMS showed that the reaction was not completed and 0.2 additional equivalents of MeI were added and the reaction stirred for additional 4 hours. The solvent was evaporated and to the crude, water was added, and the sample was freeze-dried to give [(2S)-2-[(3,5-difluorophenyl)pyrrolidin-1-yl]-(3-methylimidazol)-3-ium-1-yl-methanone;iodide **IIa** (614 mg, 93%, 95 % purity) as a white solid.

^{1}H NMR (400.23 MHz, DMSO-d6) δ ppm 9.73 (br s, 1 H), 8.20 (br s, 1 H), 7.85 (br s, 1 H), 7.22 (br d, *J*=6.85 Hz, 2 H), 7.12 (t, *J*=9.02, 9.02 Hz, 1 H), 5.10 (t, *J*=7.46, 7.46 Hz, 1 H), 3.99 (m, 1 H), 3.92 (s, 3H), 3.77 (m, 1H), 2.45 (m, 1H), 1.93 (m, 2H), 1.76 (m, 1H). LCMS (Method1): RT 1.01 min, 292 $[M]^+$.

Synthesis of Int IIIa

To a solution of IIa (360 mg, 0.086 mmol) in DMF (3.3 mL), (S)-3-Hydroxypyrrolidine (82 mg, 1.1 eq) and Et_3N (0.3 mL, 2.5 eq) were added. The mixture was stirred at RT overnight. The mixture was then concentrated in vacuo, diluted in CH_2Cl_2, and washed with a citric acid solution (10%). The organic layer was then dried over Na_2SO_4, filtered and evaporated to give intermediate **IIIa** as a white solid (200mg, 79%, 94% purity) which was used in the next step without further purification.

^{1}H NMR (400.23 MHz, DMSO-d6) δ ppm 7.95 (s, 1 H), 7.01 (tt, *J*=9.32, 2.35, Hz, 1 H), 6.95 (d, 7.74 Hz, 2 H), 4.92 (m, 2 H), 4.17 (m,1 H), 3.57 (m, 1 H), 3.46 (m, 1H), 3.43-3.21 (m, 2 H), 3.16 (m, 1H), 2.89 (s, 2 H), 2.73 (s, 2 H), 2.25 (m, 1 H), 1.84 (m, 3 H), 1.62 (m, 2 H). LCMS (Method 1): RT 1.61 min, 297.0 $[M+H]^+$.

Synthesis of Intermediate IIb

Step 1b:

To a solution of (2S)-2-(3,4difluorophenyl)pyrrolidine hydrochloride (250 mg, 1.36 mmol) in CH_2Cl_2 (1.72 mL), CDI (202.1 mg, 1.1 eq) and Et3N (0.32 mL, 2 eq) were added, and the reaction was stirred at rt for 2h. Then the mixture was washed with water and the water layer with CH_2Cl_2 (2 times). The combined organic layers were dried over Na2SO4, filtered and evaporated to yield (2S)-2-[(3,4-difluorophenyl)pyrrolidine-1-yl]-imidazol-1-yl-methanone (290 mg, 76%) which was used in the following step without further purification. LCMS Method 1: RT 1.34 min, 278.1 [M+H]+.

Step 2b:

To a solution of (2S)-2-[(3,4-difluorophenyl)pyrrolidin-1-yl]-imidazol-1-yl-methanone (290 mg, 1.04 mmol) in acetonitrile (2.1 mL), iodomethane (0.013 mL, 1.5 eq) was added and the mixture was stirred at RT overnight. LCMS showed that the reaction was not completed and 0.2 additional equivalents of MeI were added and the reaction stirred for another night. The solvent was evaporated and the crude product [(2S)-2-[(3,4-difluorophenyl)pyrrolidin-1-yl]-(3-methylimidazol)-3-ium-1-yl-methanone;iodide **IIb** (540 mg, 91% purity) was used as such without further purification. LCMS Method 1: RT 1.01 min, 292 [M]+.

Synthesis of Int IIIb

To a solution of IIb (400 mg, 0.95 mmol) in DMF (3.7 mL), (S)-3-hydroxypyrrolidine (91.4 mg, 1.1 eq) and Et_3N (0.3 mL, 2.5 eq) were added. The mixture was stirred at rt for 2h. The mixture was then concentrated in vacuo, diluted in CH_2Cl_2, and washed with a citric acid solution (10%) and twice with brine. The organic layer was then dried over Na_2SO_4, filtered and evaporated to give intermediate **IIIb** (200mg, 71%, 80% purity) as a white solid which was used in

the next step without further purification. LCMS Method 1: RT 1.60 min, 297.1 [M+H]+.

Synthesis of Compound 1:

To a solution of ((S)-2-(3,5-difluorophenyl)pyrrolidin-1-yl) ((S)-3-hydroxypyrrolidin-1-yl)methanone (IIIa) (50 mg, 0.17 mmol) in THF (3.4 mL), 3-cyanophenol (22.1 mg, 1.1 eq) and triphenylphosphine resin (178 mg, 1.5 eq) were added. The reaction was stirred for 5 min and diisopropyl-azodicarboxylate (0.043 mL, 1.3 eq) was added. The mixture was stirred at 50°C overnight and then it was filtered, washed with THF and the solvent was evaporated. The residue was dissolved in DMF and the product purified by HPLC to give Cpd 1 (24.9 mg, 37%) as a white solid.

^{1}H NMR (700.13 MHz, DMSO-d6, δ ppm): 7.49 (dd, *J* = 8.4, 7.5 Hz, 1H), 7.44 – 7.42 (m, 1H), 7.41 (dt, *J* = 7.6, 1.2 Hz, 1H), 7.26 (ddd, *J* = 8.5, 2.6, 1.0 Hz, 1H), 6.99 (tt, 1H), 6.93 – 6.87 (m, 2H), 5.13 – 5.08 (m, 1H), 4.91 (dd, *J* = 8.9, 7.0 Hz, 1H), 3.70 (dd, *J* = 12.4, 4.0 Hz, 1H), 3.62 (td, *J* = 9.6, 6.4 Hz, 1H), 3.54 (td, *J* = 10.4, 7.3 Hz, 1H), 3.46 (ddd, *J* = 10.1, 8.0, 2.7 Hz, 1H), 3.41 – 3.36 (m, 2H), 2.32 – 2.25 (m, 1H), 2.12 – 2.03 (m, 2H), 1.87 (dtt, *J* = 12.4, 6.3, 3.0 Hz, 1H), 1.82 – 1.73 (m, 1H), 1.56 (dddd, *J* = 12.2, 10.8, 9.0, 6.4 Hz, 1H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 162.29 (dd, *J* = 245.6, 13.2 Hz, 2 C), 159.91, 156.98, 150.05 (t, *J* = 8.4 Hz), 130.94, 124.74, 121.19, 118.56, 118.26, 112.35, 108.62 (dd, signal of higher order, 2 C), 101.61 (t, *J* = 25.9 Hz), 76.35, 61.08 (d, *J* = 2.1 Hz), 53.05, 49.75, 45.29, 34.88, 30.20, 25.32.

^{19}F NMR (376 MHz, DMSO-d6, δ ppm): -110.18 – -110.30 (m).

HRMS (m/z): [M]+ calc. for $C_{22}H_{21}F2N_3O_2$: 397.1602; Found 398.1669 [M+H]+.

Synthesis of Compound 3:

Cpd 3 was synthesized in an equivalent manner to **1**, but using 4-Fluorophenol (20.8 mg, 1.1 eq) to give **3** as a white solid (27.5 mg, 42%).

^{1}H NMR (700.13 MHz, DMSO-d6, δ ppm): 7.13 – 7.09 (m, 2H), 7.03 – 6.98 (m, 1H), 6.95 – 6.89 (m, 4H), 4.97 – 4.94 (m, 1H), 4.91 (dd, *J* = 9.0, 7.0 Hz, 1H), 3.65 (dd, *J* = 12.2, 4.0 Hz, 1H), 3.63 – 3.58 (m, 1H), 3.56 – 3.50 (m, 1H), 3.47 – 3.43 (m, 1H), 3.39 – 3.34 (m, 2H), 2.31 – 2.26 (m, 1H), 2.03 (ddd, *J* = 10.6, 6.7, 3.5 Hz, 2H), 1.90 – 1.84 (m, 1H), 1.81 – 1.74 (m, 1H), 1.60 – 1.53 (m, 1H).

^{13}C NMR (176 MHz, DMSO) δ 162.30 (dd, *J* = 245.5, 13.1 Hz, 2 C), 159.95, 156.66 (d, *J* = 236.4 Hz), 153.04 (d, *J* = 1.9 Hz), 150.08 (t, *J* = 8.4 Hz), 117.06 (d, *J* = 8.1 Hz), 115.92 (d, *J* = 22.9 Hz), 108.64 (dd, signal of higher order, 2 C), 101.62 (t, *J* = 25.9 Hz), 76.39, 61.09 (t, *J* = 2.0 Hz), 53.06, 49.77, 45.31, 34.89, 30.20, 25.32.

^{19}F NMR (376 MHz, DMSO-d6, δ ppm): -110.19 – -110.30 (m), -123.40 – -123.55 (m).

HRMS (m/z): $[M]^+$ calc. for $C_{21}H_{21}F_3N_2O_2$: 390.1555; Found 391.1627 $[M+H]^+$.

Synthesis of Compound 4:

To a solution of ((S)-2-(3,4-difluorophenyl)pyrrolidin-1-yl)((S)-3-hydroxypyrrolidin-1-yl)methanone (IIIb) (50 mg, 0.17 mmol) in THF (3.4 mL), 3-Chloro-4-fluorophenol (27.2 mg, 1.1 eq) and triphenylphosphine resin (178 mg, 1.5 eq) were added. The reaction was stirred for 5 min and diisopropyl-azodicarboxylate (0.043 mL, 1.3 eq) were added. The mixture was stirred at 50°C overnight. The mixture was filtered, washed with THF and the solvent was evaporated. The residue was dissolved in DMF and the product purified by HPLC to give **Cpd 4** (28.2mg, 39%) as a white solid.

^{1}H NMR (700.13 MHz, DMSO-d6, δ ppm): 7.34 (t, *J* = 9.1 Hz, 1H), 7.28 (dt, *J* = 10.7, 8.4 Hz, 1H), 7.22 (ddd, *J* = 12.0, 7.8, 2.2 Hz, 1H), 7.17 (dd, *J* = 6.1, 3.0 Hz, 1H), 7.05 – 7.02 (m, 1H), 6.94 – 6.91 (m, 1H), 5.03 – 5.00 (m, 1H), 4.88 (dd, *J* = 8.9, 7.0 Hz, 1H), 3.65 (dd, *J* = 12.3, 4.0 Hz, 1H), 3.60 (td, *J* = 9.6, 6.4 Hz, 1H), 3.54 – 3.49 (m, 1H), 3.47 – 3.43 (m, 1H), 3.37 – 3.32 (m, 2H), 2.29 – 2.24 (m, 1H), 2.04 – 2.01 (m, 2H), 1.89 – 1.84 (m, 1H), 1.81 – 1.73 (m, 1H), 1.60 – 1.54 (m, 1H).

^{13}C NMR (176 MHz, DMSO) δ 159.92, 153.42 (d, *J* = 2.3 Hz), 151.96 (d, *J* = 239.3 Hz), 148.58 (dd, *J* = 473.4, 12.6 Hz), 149.00 – 148.23 (m), 143.03 – 142.64 (m), 122.26 (dd, *J* = 6.3, 3.1 Hz), 119.81 (d, *J* = 18.9 Hz), 117.36 – 117.15 (m), 116.90 (d, *J* = 16.9 Hz), 116.08 (d, *J* = 6.8 Hz), 114.49 (d, *J* = 17.2 Hz), 76.75, 60.70, 52.93, 49.75, 45.27, 35.14, 30.18, 25.26.

^{19}F NMR (376 MHz, DMSO-d6, δ ppm): -126.38 – -127.00 (m), -139.04 – -139.19 (m), -142.13 – -142.30 (m).

HRMS (m/z): $[M]^+$ calc. for $C_{21}H_{20}ClF_3N_2O_2$: 424.1165; Found 425.1123 $[M+H]^+$.

Synthesis of Compound 6:

Cpd 6 was synthesized in an equivalent manner to Compound 4 but using 4-(1-Methyl-1H-Pyrazol-4-yl)phenol (32.3 mg, 1.1 eq) to give **Compound 6** as a white solid (31.1 mg, 41%).

^{1}H NMR (700.13 MHz, DMSO-d6, δ ppm): 8.02 (d, *J* = 0.8 Hz, 1H), 7.77 (d, *J* = 0.8 Hz, 1H), 7.50 – 7.44 (m, 2H), 7.30 – 7.21 (m, 2H), 7.06 – 7.00 (m, 1H), 6.92 – 6.88 (m, 2H), 4.99 (td, *J* = 4.0, 1.6 Hz, 1H), 4.88 (dd, *J* = 8.9, 7.0 Hz, 1H), 3.84 (s, 3H), 3.67 (dd, *J* = 12.1, 4.1 Hz, 1H), 3.63 – 3.57 (m, 1H), 3.57 – 3.50 (m, 1H), 3.48 – 3.43 (m, 1H), 3.39 – 3.32 (m, 2H), 2.27 (dtd, *J* = 12.8, 6.7, 3.2 Hz, 1H), 2.08 – 1.99 (m, 2H), 1.87 (dtt, *J* = 12.3, 6.2, 2.9 Hz, 1H), 1.82 – 1.73 (m, 1H), 1.56 (dddd, *J* = 12.2, 10.8, 9.0, 6.3 Hz, 1H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 159.97, 155.08, 148.58 (dd, *J* = 474.4, 12.7 Hz), 148.58 (t, *J* = 13.2 Hz), 142.85 (t, *J* = 4.1 Hz), 135.58, 127.08, 126.18 (2 C), 125.59, 122.33 – 122.07 (m), 121.60, 116.91 (d, *J* = 16.9 Hz), 116.07 (2 C), 114.53 (d, *J* = 17.2 Hz), 75.78, 60.73, 53.10, 49.79, 45.37, 38.56, 35.13, 30.29, 25.28.

^{19}F NMR (376 MHz, DMSO-d6, δ ppm): -139.03 – -139.19 (m), -142.10 – -142.30 (m).

HRMS (m/z): $[M]^+$ calc. for $C_{25}H_{26}F_2N_4O_2$: 452.2024. Found 453.2102 $[M+H]^+$.

Synthesis of Compound 2:

To a solution of IIa (25 mg, 0.059 mmol) in DMF (0.23 mL), 3-(3-chloro-4-fluorophenoxy)-azetidine hydrochloride (19.1 mg, 1.1 eq) and Et_3N (0.029 mL, 3.5 eq) were added. The mixture was stirred at RT overnight. The mixture was then concentrated in vacuo, diluted in CH_2Cl_2, and washed with a citric acid solution (10%). The organic layer was then dried over Na_2SO_4, filtered, and evaporated. The crude product was purified by HPLC to give **Cpd 2** as a white solid (16.2 mg, 66%).

^{1}H NMR (700.13 MHz, DMSO-d6, δ ppm): 7.35 (dd, *J* = 9.1, 9.1 Hz, 1H), 7.07 (dd, *J* = 6.0, 3.1 Hz, 1H), 7.05 – 7.01 (m, 1H), 6.98 – 6.94 (m, 2H), 6.88 – 6.85 (m, 1H), 5.03 – 4.99 (m, 1H), 4.91 (dd, *J* = 7.6, 5.7 Hz, 1H), 4.38 (dd, *J* = 9.6, 6.5 Hz, 1H), 4.29 (dd, *J* = 9.6, 6.4 Hz, 1H), 3.87 (dd, *J* = 9.7, 4.0 Hz, 1H), 3.82 – 3.79 (m, 1H), 3.57 – 3.52 (m, 1H), 3.42 – 3.38 (m, 1H), 2.26 – 2.20 (m, 1H), 1.83 – 1.78 (m, 2H), 1.64 – 1.59 (m, 1H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 162.31 (dd, *J* = 245.7, 13.1 Hz, 2 C), 160.00, 152.98 (d, *J* = 2.2 Hz), 152.89, 151.53, 149.84 (t, *J* = 8.3 Hz), 119.98 (d, *J* = 19.1 Hz), 117.45 (d, *J* = 22.4 Hz), 116.15, 115.11 (d, *J* = 6.9 Hz), 108.66 (dd, signal of higher order, 2 C), 101.72 (t, *J* = 25.9 Hz), 66.63, 60.58 (d, *J* = 2.3 Hz), 56.78, 56.69, 47.59, 34.42, 24.09.

^{19}F NMR (376 MHz, DMSO-d6, δ ppm): -110.06 – -110.19 (m), -126.31 – -126.44 (m).

HRMS (m/z): [M]+ calc. for $C_{20}H_{18}ClF_3N_2O_2$: 410.1009; Found 411.1069 [M+H]+.

Synthesis of Compound 5:

To a solution of IIa (25 mg, 0.059 mmol) in DMF (0.23 mL), 3-(3,4-dichlorophenoxy)-azetidine hydrochloride (16 mg, 1.1 eq) and Et_3N (0.029 mL, 3.5 eq) were added. The mixture was stirred at RT overnight. The mixture was then concentrated in vacuo, diluted in CH_2Cl_2, and washed with a citric acid solution (10%). The organic layer was then dried over Na_2SO_4, filtered, and evaporated. The crude product was purified by HPLC to give **Cpd 5** as a white solid (15.8 mg, 62%).

^{1}H NMR (700.13 MHz, DMSO-d6, δ ppm): 7.54 (d, *J* = 8.9 Hz, 1H), 7.13 (d, *J* = 2.9 Hz, 1H), 7.03 (tt, *J* = 9.3, 2.4 Hz, 1H), 6.98 – 6.93 (m, 2H), 6.89 (dd, *J* = 8.9, 2.9 Hz, 1H), 5.06 – 5.01 (m, 1H), 4.91 (dd, *J* = 7.6, 5.6 Hz, 1H), 4.38 (dd, *J* = 9.6, 6.5 Hz, 1H), 4.32 – 4.27 (m, 1H), 3.88 (dd, *J* = 9.7, 4.0 Hz, 1H), 3.84 – 3.79 (m, 1H), 3.57 – 3.52 (m, 1H), 3.43 – 3.37 (m, 1H), 2.26 – 2.19 (m, 1H), 1.84 – 1.77 (m, 2H), 1.65 – 1.58 (m, 1H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 163.01 (d, *J* = 13.1 Hz), 161.62 (d, *J* = 13.2 Hz), 159.97, 155.83, 149.83 (t, *J* = 8.3 Hz), 131.82, 131.20, 123.21, 116.54, 115.58, 108.66 (dd, *J* = 20.6, 4.3 Hz), 101.73 (t, *J* = 25.9 Hz), 66.63, 60.58 (d, *J* = 2.3 Hz), 56.69 (d, *J* = 22.1 Hz), 47.59, 34.42, 24.09.

^{19}F NMR (376 MHz, DMSO-d6, δ ppm): δ -110.06 – -110.19 (m).

HRMS (m/z): $[M]^+$ calc. for $C_{20}H_{18}Cl_2F_2N_2O_2$: 426.0713; Found 427.0782 $[M+H]^+$.

Synthesis of 1-(benzo[d][1,3]dioxol-5-yl)-3-(4-chlorobenzyl)-1,3-dihydro-2H-imidazol-2-one (7): To a solution of commercially available 1-(1,3-benzodioxol-5-yl)-3H-imidazol-2-one (IV) (100 mg, 0.49 mmol) in DMF (2 mL) the p-chloro benzyl bromide was added (111 mg, 0.54 mmol, 1.10 eq.) and K_2CO_3 (1.50 eq.) in sequence. The resulting mixture was stirred at 80 °C for 16 h. The reaction was diluted with ice-cooled NH_4Cl solution (sat. aq., 6 mL) and water (10 mL), and then extracted with EtOAc (3 × 15 mL). The combined organic layers were washed with brine (20 mL), dried over anhydrous Na_2SO_4 solid, filtered and concentrated under reduced pressure to give a residue. The residue was purified by column chromatography (SiO_2, PE/EtOAc = 10/1 to 1/1) to afford the corresponding product **Cpd 7** (89.13 mg, 99.9% purity, 55.4% yield) was obtained as a white solid.

^{1}H NMR (700 MHz, DMSO-d6, δ ppm): 7.44 – 7.41 (m, 2H), 7.33 – 7.31 (m, 3H), 7.13 (dd, *J* = 8.4, 2.2 Hz, 1H), 6.99 – 6.94 (m, 2H), 6.77 (d, *J* = 3.1 Hz, 1H), 6.05 (s, 2H), 4.78 (s, 2H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 151.06, 147.44, 144.62, 136.67, 132.06, 131.51, 129.39, 128.53, 113.99, 112.06, 109.95, 108.09, 102.85, 101.40, 45.55.

HRMS (m/z): $[M]^+$ calc. for $C_{17}H_{13}ClN_2O_3$: 328.0615; Found 328.0612 $[M+H]^+$.

1-(benzo[d][1,3]dioxol-5-yl)-3-(3,4-difluorobenzyl)-1,3-dihydro-2H-imidazol-2-one (9) was prepared in the same way but using 3-4-diF-benzylbromide (112 mg, 0.539 mmol, 1.10 eq.) to yield Cpd 9 (74.11 mg, 99.9% purity, 0.224 mmol, 45.8% yield) was obtained as a white solid.

^{1}H NMR (400 MHz, DMSO-d6, δ ppm): 7.46 - 7.36 (m, 2H), 7.32 (d, J = 2.0 Hz, 1H), 7.18 - 7.10 (m, 2H), 7.00 - 6.94 (m, 2H), 6.80 (d, J = 3.2 Hz, 1H), 6.05 (s, 2H), 4.77 (s, 2H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 151.03, 149.72 (dd, *J* = 81.9, 12.7 Hz), 148.32 (dd, *J* = 81.1, 12.6 Hz), 147.44, 144.66, 135.37 (dd, *J* = 5.5, 3.7 Hz), 131.47, 124.45 (dd, *J* = 6.6, 3.4 Hz), 117.66 (d, *J* = 17.2 Hz), 116.73 (d, *J* = 17.3 Hz), 114.06, 111.97, 110.07, 108.08, 102.91, 101.41, 45.26.

^{19}F NMR (376 MHz, DMSO-d6, δ ppm): -138.26 – -138.44 (m), -140.34 – -140.53 (m).

HRMS (m/z): $[M]^+$ calc. for $C_{17}H_{12}F_2N_2O_3$: 330.0816; Found 330.0809 $[M+H]^+$.

Synthesis of 1-(4-chlorobenzyl)-3-(6-(((2-fluoropyridin-4-yl)methyl)amino)pyrimidin-4-yl)-1,3-dihydro-2H-imidazol-2-one (8):

Step 1: To a solution of 1-[(4-chlorophenyl)methyl]-2,3-dihydro-1H-imidazol-2-one (Va) (800 mg, 3.83 mmol, 1.00 eq.) in DMF (8 mL) was added NaH (60% purity, 0.184 g, 4.60 mmol, 1.20 eq.) at 0 ℃, After stirring at 25 ℃ for 30 min, 4,6-dichloropyrimidine (1.20 g, 8.05 mmol, 2.10 eq.) was added at 0 ℃, and the resulting mixture was stirred at 25 ℃ for 2 h. The reaction mixture was diluted with ice-water (30 mL) and then extracted with EtOAc (3 × 20 mL). The combined organic layers were washed with brine (30 mL), dried with anhydrous Na2SO4 solid, filtered and concentrated under reduced pressure to give a residue. The

residue was purified by column chromatography (SiO2, PE/EtOAc = 25/1 to 4/1) to afford 1-(4-chlorobenzyl)-3-(6-chloropyrimidin-4-yl)-1,3-dihydro-2H-imidazol-2-one (660 mg, 2.06 mmol, 53.6% yield) as a yellow solid.

^{1}H NMR (400 MHz, DMSO-d6) δ = 8.90 (s, 1H), 8.45 (s, 1H), 7.47 - 7.41 (m, 2H), 7.39 - 7.32 (m, 3H), 7.00 (d, J = 3.2 Hz, 1H), 4.84 (s, 2H).

Step 2: To a solution of (2-fluoropyridin-4-yl)methanamine hydrochloride (200 mg, 0.623 mmol, 1.00 eq.) in DMA (3 mL) was added1-(4-chlorobenzyl)-3-(6-chloropyrimidin-4-yl)-1,3-dihydro-2H-imidazol-2-one (VI) (203 mg, 1.25 mmol, 2.00 eq.) and DIEA (0.320 mL, 1.87 mmol, 3.00 eq.), and the resulting mixture was stirred at 80 ℃ for 16 h. The mixture was diluted with water (10 mL) and then extracted with EtOAc (2 × 20 mL). The combined organic layers were washed with brine (20 mL), dried over anhydrous Na_2SO_4 solid, filtered and concentrated under reduced pressure to give a residue. The residue was purified by column chromatography (SiO2, PE/EtOAc = 10/1 to 1/1) to afford **Cpd 8** (70.12 mg, 27.4% yield) as an off-white solid.

^{1}H NMR (700 MHz, DMSO-d6, δ ppm): 8.31 (d, *J* = 1.1 Hz, 1H), 8.27 (s, 1H), 8.16 (d, *J* = 5.2 Hz, 1H), 7.63 (s, 1H), 7.45 – 7.40 (m, 2H), 7.32 (d, *J* = 8.1 Hz, 2H), 7.27 (dd, *J* = 12.3, 3.9 Hz, 2H), 7.04 (s, 1H), 6.83 (d, *J* = 3.2 Hz, 1H), 4.81 (s, 2H), 4.64 (s, 2H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 164.02, 163.39, 162.69, 157.94, 155.93, 153.32, 150.93, 147.50, 147.41, 136.29, 132.16, 129.35, 128.57, 120.28 (d, *J* = 3.7 Hz), 113.21, 107.25, 107.04, 106.62, 91.07, 45.41, 42.43.

^{19}F NMR (376 MHz, DMSO) δ -69.04 – -69.70 (m).

HRMS (m/z): $[M]^+$ calc. for $C_{20}H_{16}ClFN_6O$: 410.1058; Found 410.1055 $[M+H]^{+\cdot}$

Synthesis of 1-benzyl-3-(4-(morpholinosulfonyl)phenyl)-1,3-dihydro-2H-imidazol-2-one (10):

To a solution of commercially available 1-benzyl-2,3-dihydro-1H-imidazol-2-one (200 mg, 1.15 mmol) and 4-((4-bromophenyl)sulfonyl)morpholine (422 mg, 1.38 mmol, 1.20 eq.) in toluene (5 mL) was added CuI (10 mol%), DMEDA (20 mol%) and K_2CO_3 (2.50 eq.). The resulting mixture was stirred at 100 °C for 16 h under N2 atmosphere. The reaction was diluted with ice-cooled NH4Cl solution (sat. aq., 10 mL) and water (10 mL), and then extracted with EtOAc (3 × 15 mL). The combined organic layers were washed with brine (20 mL), dried over anhydrous Na_2SO_4 solid, filtered and concentrated under reduced pressure to give a residue. The residue was purified by column chromatography to afford **Cpd 10** (23.67 mg, 99.9% purity, 51.6% yield) as an off white solid.

^{1}H NMR (700 MHz, DMSO-d6, δ ppm): 8.12 – 8.08 (m, 2H), 7.82 – 7.78 (m, 2H), 7.38 – 7.35 (m, 2H), 7.32 – 7.28 (m, 3H), 7.26 (d, *J* = 3.2 Hz, 1H), 6.92 (d, *J* = 3.2 Hz, 1H), 4.82 (s, 2H), 3.65 – 3.61 (m, 4H), 2.90 – 2.86 (m, 4H).

^{13}C NMR (176 MHz, DMSO-d6, δ ppm): 151.07, 141.21, 137.33, 130.03, 128.92, 128.60, 127.51, 127.46, 119.91, 113.86, 108.54, 65.23, 46.27, 45.86.

HRMS (m/z): [M]+ calc. for $C_{20}H_{21}N_3O_4S$: 399.1253; Found 399.1251 $[M+H]^+$.

Cathepsin A

Analytical UPLC method

Analytical HPLC was performed on a Waters Acquity UPLC system with a Waters Acquity BEH 1.7 µm C18 column (2.1 x 50 mm) at 55 °C with a gradient elution at a flow rate of 1.0 mL/min and UV monitoring at 220 nm or 254 nm. Mass determination was performed by coupled Waters single quadrupole mass spectrometer SQD2 by electrospray ionization and detection of positive and negative ions in parallel (ESI+/-) over a mass range of 100-2000 m/z.

The gradients were set up as 2% B, after an initial wait period of 0.2 min to 98% B over 3.6 min and then 98% B for 0.5 min coming back to 2% B within 0.2 min. Buffer A = 0.05 % formic acid in water and B = 0.035 % formic acid in acetonitrile.

column: Waters ACQUITY UPLC® BEH™ C18 1.7 µm (50 x 2.1 mm) at 55 °C

solvent: H_2O+0.05%FA : ACN+0.035%FA (flow 1.0 ml/min)

gradient: 98:2 (0 min) to 98:2 (0.2min) to 2:98 (3.8 min) to 2:98 (4.3 min) to 98:2 (4.5 min)

NMR method

NMR spectra were recorded on a Bruker AVANCE 400 spectrometer operating at a proton frequency of 400.32 MHz. The instrument was equipped with 5 mm BBI room temperature probe. For assignment of proton resonances 1D-1H spectra were acquired. All experiments were conducted using samples of compound dissolved in d6-DMSO at 300 K. Chemical shifts were referenced to the internal standard Tetramethylsilane (TMS) (1H: 0.0 ppm).

Instrument: 400 NMR

Frequency: 400.32 MHz

Console: AVANCE II 400

Magnet: Ultrashield 400 plus

Probe: 5mm PABBI 1H/1D-BB with z-gradient

List of individual compounds:

3-[[1-(3-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-(3-methylphenyl)propanoic acid (11)

^{1}H NMR (400 Mhz, DMSO) δ 12.26 (s, 1H), 12.24 (br s, 1H), 8.59 (d, J = 8.8 Hz, 1H), 7.75 – 7.67 (m, 2H), 7.59 – 7.50 (m, 1H), 7.23 – 7.15 (m, 4H), 7.07 – 7.02 (m, 1H), 5.87 (s, 1H), 5.42 – 5.33 (m, 1H), 2.95 (dd, J = 15.9, 8.2 Hz, 1H), 2.80 (dd, J = 15.9, 6.2 Hz, 1H), 2.29 (s, 3H).

^{13}C NMR (125 MHz, DMSO) δ 171.92, 162.05 (d: J = 243.0 Hz), 161.14 (d: J = 242.6 Hz), 160.52, 154.18, 146.01, 139.81 (d: J = 10.7 Hz), 138.78 (d: J = 3.0 Hz), 130.76 (d: J = 9.2 Hz), 128.64 (d: J = 8.1 Hz, 2 C), 117.14 (d: J = 2.7 Hz), 114.92 (d: J = 21.2 Hz, 2 C), 113.03 (d: J = 21.0 Hz), 108.51 (d: J = 26.2 Hz), 87.54, 48.49, 40.08.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_4$: 383.1276; Found 383.1267 $[M+H]^+$.

3-[[1-(2-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-(3-methylphenyl)propanoic acid (12)

^{1}H NMR (400 Mhz, DMSO) δ 12.18 (br s, 1H), 11.74 (br s, 1H), 8.51 (d, J = 8.8 Hz, 1H), 7.59 – 7.51 (m, 2H), 7.47 – 7.40 (m, 1H), 7.38 – 7.32 (m, 1H), 7.22 – 7.16 (m, 3H), 7.06 – 7.01 (m, 1H), 5.83 (s, 1H), 5.40 – 5.31 (m, 1H), 2.93 (dd, J = 15.9, 8.1 Hz, 1H), 2.76 (dd, J = 15.9, 6.3 Hz, 1H), 2.28 (s, 3H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_4$: 383.1276; Found 383.1283 $[M+H]^+$.

3-[[1-(4-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-(3-methylphenyl)propanoic acid (13)

^{1}H NMR (400 Mhz, DMSO) δ 12.22 (br s, 1H), 12.00 (s, 1H), 8.54 (d, J = 8.8 Hz, 1H), 7.83 – 7.76 (m, 2H), 7.39 – 7.31 (m, 2H), 7.23 – 7.17 (m, 3H), 7.07 – 7.02 (m, 1H), 5.86 (s, 1H), 5.41 – 5.32 (m, 1H), 2.94 (dd, J = 15.9, 8.2 Hz, 1H), 2.79 (dd, J = 15.9, 6.2 Hz, 1H), 2.28 (s, 3H).

^{13}C NMR (176 MHz, DMSO) δ 172.04, 160.58, 160.32, (d: J = 243.7 Hz), 153.64, 145.75, 142.54, 137.19, 134.68 (d: J = 2.6 Hz), 128.13, 127.56, 127.28, 124.14 (d: J = 8.5 Hz, 2 C), 123.70, 115.70 (d: J = 22.8 Hz, 2 C), 87.25, 49.00, 40.08, 21.08.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_4$: 383.1276; Found 383.1269 $[M+H]^+$.

3-(4-chlorophenyl)-3-[[1-(4-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (14)

^{1}H NMR (400 Mhz, DMSO) δ 12.29 (br s, 1H), 12.01 (s, 1H), 8.64 (d, J = 8.7 Hz, 1H), 7.83 – 7.75 (m, 2H), 7.46 – 7.30 (m, 6H), 5.85 (s, 1H), 5.42 – 5.33 (m, 1H), 2.96 (dd, J = 16.0, 8.2 Hz, 1H), 2.81 (dd, J = 16.0, 6.4 Hz, 1H).

^{13}C NMR (176 MHz, DMSO) δ 171.88, 160.71, 160.35 (d: J = 243.8 Hz), 153.61, 145.62, 141.63, 134.63 (d: J = 2.6 Hz), 131.49, 128.61 (2 C), 128.16 (2 C), 124.21 (d: J = 8.5 Hz, 2C), 115.71 (d: J = 22.8 Hz, 2 C), 87.27, 48.55, 40.00.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}ClFN_3O_4$: 403.0730; Found 403.0734 $[M+H]^+$.

3-(4-fluorophenyl)-3-[[1-(4-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (15)

^{1}H NMR (400 Mhz, DMSO) δ 12.26 (br s, 1H), 12.00 (s, 1H), 8.62 (d, J = 8.8 Hz, 1H), 7.83 – 7.75 (m, 2H), 7.48 – 7.41 (m, 2H), 7.39 – 7.30 (m, 2H), 7.18 – 7.10 (m, 2H), 5.85 (s, 1H), 5.44 – 5.35 (m, 1H), 2.96 (dd, J = 16.0, 8.2 Hz, 1H), 2.80 (dd, J = 16.0, 6.4 Hz, 1H).

^{13}C NMR (176 MHz, DMSO) δ 171.94, 161.13 (d: J = 242.5 Hz), 160.44, 160.34 (d: J = 243.8 Hz), 153.64, 145.68, 138.82 (d: J = 2.9), 134.65 (d: J = 2.6 Hz), 128.66 (d: J = 8.1 Hz, 2 C), 124.20 (d: J = 8.5 Hz, 2 C), 115.71 (d: J = 22.8 Hz, 2 C), 114.90 (d: J = 21.2 Hz, 2 C), 87.25, 48.47, 40.06.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}F_2N_3O_4$: 387.1025; Found 387.1019 $[M+H]^+$.

3-(4-fluorophenyl)-3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (16)

^{1}H NMR (400 Mhz, DMSO) δ 12.28 (br s, 1H), 11.97 (br s, 1H), 8.61 (d, J = 8.6 Hz, 1H), 7.50 – 7.31 (m, 5H), 7.19 – 7.09 (m, 2H), 6.97 – 6.89 (m, 1H), 5.85 (s, 1H), 5.44 – 5.35 (m, 1H), 3.80 (s, 3H), 2.94 (dd, J = 15.9, 7.7 Hz, 1H), 2.81 (dd, J = 15.9, 5.4 Hz, 1H).

^{13}C NMR (176 MHz, DMSO) δ 171.96, 161.13 (d: J = 242.5 Hz), 160.68, 159.52, 153.71, 145.57, 139.30, 138.80 (d: J = 2.9 Hz), 129.78, 128.65 (d: J = 8.1 Hz, 2 C), 114.90 (d: J = 21.2 Hz, 2 C), 114.16, 111.80, 108.13, 87.45, 55.34, 48.46, 40.04.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_5$: 399.1225; Found 399.1227 $[M+H]^+$.

3-(4-fluorophenyl)-3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (17)

^{1}H NMR (400 Mhz, DMSO) δ 12.27 (br s, 1H), 11.74 (br s, 1H), 8.56 (d, J = 8.8 Hz, 1H), 7.65 – 7.59 (m, 2H), 7.48 – 7.40 (m, 2H), 7.18 – 7.10 (m, 2H), 7.08 – 7.01 (m, 2H), 5.83 (s, 1H), 5.44 – 5.34 (m, 1H), 3.80 (s, 3H), 2.96 (dd, J = 16.0, 8.1 Hz, 1H), 2.80 (dd, J = 16.0, 6.2 Hz, 1H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_5$: 399.1225; Found 399.1233 $[M+H]^+$.

3-(4-tert-butylphenyl)-3-[[1-(4-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (18)

1H NMR (400 Mhz, DMSO) δ 12.21 (br s, 1H), 12.00 (br s, 1H), 8.54 (d, J = 8.8 Hz, 1H), 7.83 – 7.76 (m, 2H), 7.39 – 7.29 (m, 6H), 5.85 (s, 1H), 5.42 – 5.33 (m, 1H), 2.94 (dd, J = 15.9, 8.1 Hz, 1H), 2.80 (dd, J = 15.9, 6.3 Hz, 1H), 1.25 (s, 9H).

^{13}C NMR (125 MHz, DMSO) δ 172.09, 160.57, 160.33 (d: J = 243.7 Hz), 153.54, 149.21, 145.77, 139.56, 134.66 (d: J = 2.7 Hz), 126.35 (2 C), 124.95 (2 C), 124.15 (d: J = 8.5 Hz), 115.71 (d: J = 22.8 Hz), 87.26, ~39.9 (below solvent signal), 34.12, 31.11 (3 C).

HRMS (m/z): $[M+H]^+$ calc. for $C_{23}H_{24}FN_3O_4$: 425.1745; Found 425.1748 $[M+H]^+$.

3-[[1-(3-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-[4-[(2-methylpropan-2-yl)oxy]phenyl]propanoic acid (19)

^{1}H NMR (400 Mhz, DMSO) δ 12.41 – 12.05 (m, 2H), 8.60 (d, J = 8.9 Hz, 1H), 7.76 – 7.66 (m, 2H), 7.59 – 7.49 (m, 1H), 7.34 – 7.28 (m, 2H), 7.23 – 7.14 (m, 1H), 6.95 – 6.88 (m, 2H), 5.87 (s, 1H), 5.45 – 5.35 (m, 1H), 2.94 (dd, J = 15.9, 8.3 Hz, 1H), 2.80 (dd, J = 15.9, 6.2 Hz, 1H), 1.27 (s, 9H).

Compound only has a low purity,13 C data therefore not included. Compound contains 5.6:1 mixture of tBu vs OH phenol.

HRMS (m/z): $[M+H]^+$ calc. for $C_{23}H_{24}FN_3O_5$: 441.1695; Found 441.1697 $[M+H]^+$.

3-(4-tert-butylphenyl)-3-[[1-(3-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (20)

^{1}H NMR (400 Mhz, DMSO) δ 12.37 – 12.13 (m, 2H), 8.59 (d, J = 8.9 Hz, 1H), 7.75 – 7.67 (m, 2H), 7.58 – 7.50 (m, 1H), 7.36 – 7.30 (m, 4H), 7.22 – 7.15 (m, 1H), 5.86 (s, 1H), 5.43 – 5.35 (m, 1H), 2.95 (dd, J = 15.9, 8.2 Hz, 1H), 2.81 (dd, J = 15.9, 6.3 Hz, 1H), 1.25 (s, 9H).
^{13}C NMR (176 MHz, DMSO) δ 172.08, 162.06 (d: J = 243.0 Hz), 160.43, 154.06, 149.23, 146.10, 139.81 (d: J = 10.6 Hz), 139.53, 126.33 (2 C), 124.97 (2 C), 117.10 (J = 2.6 Hz), 113.02 (d: 20.9 Hz), 108.48 (d: 26.2 Hz), 87.55, 48.69, 39.99, 34.12, 31.10 (3 C).

HRMS (m/z): $[M+H]^+$ calc. for $C_{23}H_{24}FN_3O_4$: 425.1745; Found 425.1753 $[M+H]^+$.

3-(4-chlorophenyl)-3-[[1-(2-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (21)

^{1}H NMR (400 Mhz, DMSO) δ 12.28 (br s, 1H), 11.75 (br s, 1H), 8.62 (d, J = 8.7 Hz, 1H), 7.59 – 7.51 (m, 2H), 7.47 – 7.32 (m, 6H), 5.83 (s, 1H), 5.41 – 5.32 (m, 1H), 2.94 (dd, J = 16.1, 8.1 Hz, 1H), 2.78 (dd, J = 16.1, 6.4 Hz, 1H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}ClFN_3O_4$: 403.0730; Found 403.0729 $[M+H]^+$.

3-(4-chlorophenyl)-3-[[1-(3-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (22)

^{1}H NMR (400 Mhz, DMSO) δ 12.30 (br s, 1H), 12.27 (s, 1H), 8.69 (d, J = 8.7 Hz, 1H), 7.75 – 7.66 (m, 2H), 7.59 – 7.50 (m, 1H), 7.46 – 7.36 (m, 4H), 7.23 – 7.15 (m, 1H), 5.86 (s, 1H), 5.43 – 5.34 (m, 1H), 2.97 (dd, J = 16.0, 8.2 Hz, 1H), 2.82 (dd, J = 16.0, 6.3 Hz, 1H).

^{13}C NMR (176 MHz, DMSO) δ 171.87, 162.05 (d: J = 243.0 Hz), 160.58, 154.13, 145.95, 141.59, 139.78 (d: J = 10.6 Hz), 131.51, 130.77 (d: J = 9.2 Hz), 128.60 (2 C), 128.19 (2 C), 117.17 (d: J = 2.6 Hz), 113.07 (d: J = 21.0 Hz), 108.53 (d: J = 26.2 Hz), 87.56, 48.57, ~39.9 (below solvent signal).

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}ClFN_3O_4$: 403.0730; Found 403.0740 $[M+H]^+$.

3-(4-fluorophenyl)-3-[[1-(2-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (23)

^{1}H NMR (400 Mhz, DMSO) δ 12.24 (br s, 1H), 11.75 (br s, 1H), 8.59 (d, J = 8.8 Hz, 1H), 7.59 – 7.51 (m, 2H), 7.48 – 7.39 (m, 3H), 7.39 – 7.32 (m, 1H), 7.17 – 7.09 (m, 2H), 5.83 (s, 1H), 5.43 – 5.34 (m, 1H), 2.94 (dd, J = 16.0, 8.2 Hz, 1H), 2.78 (dd, J = 16.0, 6.5 Hz, 1H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}F_2N_3O_4$: 387.1025; Found 387.1014 $[M+H]^+$.

(3R)-3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-(4-methylphenyl)propanoic acid (24)

^{1}H NMR (400 Mhz, DMSO) δ 12.20 (s, 1H), 11.73 (br s, 1H), 8.47 (d, J = 8.8 Hz, 1H), 7.65 – 7.59 (m, 2H), 7.30 – 7.25 (m, 2H), 7.14 – 7.08 (m, 2H), 7.07 – 7.01 (m, 2H), 5.83 (s, 1H), 5.40 – 5.31 (m, 1H), 3.80 (s, 3H), 2.93 (dd, J = 16.1, 8.2 Hz, 1H), 2.78 (dd, J = 16.1, 6.0 Hz, 1H), 2.26 (s, 3H).

^{13}C NMR (176 MHz, DMSO) δ 172.08, 160.67, 157.87, 153.10, 145.24, 139.60, 135.95, 131.30, 128.71 (2 C), 126.56 (2 C), 123.96 (2 C), 113.99 (2 C), 87.11, 55.37, 48.72, 40.08, 20.59.

HRMS (m/z): $[M+H]^+$ calc. for $C_{21}H_{21}N_3O_5$: 395.1476; Found 395.1473 $[M+H]^+$.

3-(4-fluorophenyl)-3-[[1-(3-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (25)

^{1}H NMR (400 Mhz, DMSO) δ 12.33 – 12.21 (m, 2H), 8.66 (d, J = 8.8 Hz, 1H), 7.75 – 7.66 (m, 2H), 7.59 – 7.50 (m, 1H), 7.48 – 7.42 (m, 2H), 7.23 – 7.11 (m, 3H), 5.86 (s, 1H), 5.44 – 5.36 (m, 1H), 2.96 (dd, J = 16.0, 8.3 Hz, 1H), 2.82 (dd, J = 16.0, 6.3 Hz, 1H).

^{13}C NMR (125 MHz, DMSO) δ 171.92, 162.05 (d: J = 243.0 Hz), 161.14 (d: J = 242.6 Hz), 160.52, 154.18, 146.01, 139.81 (d: J = 10.7 Hz), 138.78 (d: J = 3.0 Hz), 130.76 (d: J = 9.2 Hz), 128.64 (d: J = 8.1 Hz, 2 C), 117.14 (d: J = 2.7 Hz), 114.92 (d: J = 21.2 Hz), 113.03 (d: J = 21.0 Hz), 108.51 (d: J = 26.2 Hz), 87.54, 48.49, 40.08.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}F_2N_3O_4$: 387.1025; Found 387.1028 $[M+H]^+$.

(3S)-3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-(4-methylphenyl)propanoic acid (26)

^{1}H NMR (400 Mhz, DMSO) δ 12.20 (br s, 1H), 11.97 (br s, 1H), 8.52 (d, J = 8.8 Hz, 1H), 7.44 – 7.33 (m, 3H), 7.32 – 7.25 (m, 2H), 7.15 – 7.08 (m, 2H), 6.96 – 6.90 (m, 1H), 5.85 (s, 1H), 5.42 – 5.32 (m, 1H), 3.81 (s, 3H), 2.94 (dd, J = 15.9, 7.9 Hz, 1H), 2.79 (dd, J = 15.9, 6.4 Hz, 1H), 2.26 (s, 3H).

^{13}C NMR (125 MHz, DMSO) δ 172.10, 160.57, 159.53, 153.63, 145.67, 139.56, 139.30, 135.99, 129.79, 128.74 (2 C), 126.55 (2 C), 114.12, 111.80, 108.08, 87.47, 55.34, 48.75, 40.09, 20.60.

HRMS (m/z): $[M+H]^+$ calc. for $C_{21}H_{21}N_3O_5$: 395.1476; Found 395.1473 $[M+H]^+$.

(3S)-3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-(4-methylphenyl)propanoic acid (27)

^{1}H NMR (400 Mhz, DMSO) δ 12.21 (br s, 1H), 11.73 (s, 1H), 8.47 (d, J = 8.9 Hz, 1H), 7.65 – 7.59 (m, 2H), 7.31 – 7.25 (m, 2H), 7.14 – 7.08 (m, 2H), 7.08 – 7.01 (m, 2H), 5.83 (s, 1H), 5.40 – 5.32 (m, 1H), 3.80 (s, 3H), 2.93 (dd, J = 15.8, 8.1 Hz, 1H), 2.78 (dd, J = 15.8, 6.3 Hz, 1H), 2.26 (s, 3H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{21}H_{21}N_3O_5$: 395.1476; Found 395.1469 $[M+H]^+$.

3-(4-chlorophenyl)-3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (28)

^{1}H NMR (400 Mhz, DMSO) δ 12.30 (br s, 1H), 11.98 (br s, 1H), 8.64 (d, J = 8.6 Hz, 1H), 7.46 – 7.33 (m, 7H), 6.97 – 6.90 (m, 1H), 5.85 (s, 1H), 5.43 – 5.34 (m, 1H), 3.81 (s, 3H), 2.97 (dd, J = 16.0, 8.1 Hz, 1H), 2.82 (dd, J = 16.0, 6.3 Hz, 1H).

^{13}C NMR (125 MHz, DMSO) δ 171.91, 160.72, 159.52, 153.64, 145.51, 141.62, 139.26, 131.49, 129.79, 128.61 (2 C), 128.17 (2 C), 114.19, 111.84, 108.16, 87.48, 55.35, 48.55, ~39.7 (below solvent signal).

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}ClN_3O_5$: 415.0930; Found 415.0922 $[M+H]^+$.

3-(4-tert-butylphenyl)-3-[[1-(2-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (29)

^{1}H NMR (400 Mhz, DMSO) δ 12.18 (s, 1H), 11.73 (s, 1H), 8.50 (d, J = 8.8 Hz, 1H), 7.59 – 7.51 (m, 2H), 7.47 – 7.40 (m, 1H), 7.38 – 7.28 (m, 5H), 5.83 (s, 1H), 5.41 – 5.32 (m, 1H), 2.93 (dd, J = 15.9, 8.1 Hz, 1H), 2.77 (dd, J = 15.9, 6.3 Hz, 1H), 1.25 (s, 9H).

^{13}C NMR (125 MHz, DMSO) δ 172.07, 160.52, 156.42 (d: J = 251.0 Hz), 154.38, 149.18, 146.93, 139.60, 130.89 (d: J = 7.7 Hz), 129.23, 126.40 (2 C), 125.24 (d: J = 12.3 Hz), 124.92 (2 C), 124.82 (d: J = 3.7 Hz), 116.43 (d: J = 19.5 Hz), 86.07, 48.68, ~40.0 (below solvent signal), 34.11, 31.10 (3 C).

HRMS (m/z): $[M+H]^+$ calc. for $C_{23}H_{24}FN_3O_4$: 425.1745; Found 425.1750 $[M+H]^+$.

3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-(3-methylphenyl)propanoic acid (30)

^{1}H NMR (400 Mhz, DMSO) δ 12.25 (br s, 1H), 11.97 (br s, 1H), 8.54 (d, J = 8.8 Hz, 1H), 7.45 – 7.33 (m, 3H), 7.24 – 7.16 (m, 3H), 7.08 – 7.01 (m, 1H), 6.97 – 6.89 (m, 1H), 5.86 (s, 1H), 5.42 – 5.32 (m, 1H), 3.81 (s, 3H), 2.95 (dd, J = 15.9, 8.2 Hz, 1H), 2.80 (dd, J = 15.9, 6.2 Hz, 1H), 2.28 (s, 3H).

^{13}C NMR (125 MHz, DMSO) δ 172.08, 160.60, 159.53, 153.63, 145.64, 142.53, 139.20, 137.21, 129.80, 128.15, 127.58, 127.27, 123.70, 114.13, 111.82, 108.10, 87.49, 55.34, 49.00, 40.08, 21.09.

HRMS (m/z): $[M+H]^+$ calc. for $C_{21}H_{21}N_3O_5$: 395.1476; Found 395.1478 $[M+H]^+$.

3-[[1-(2-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-[4-[(2-methylpropan-2-yl)oxy]phenyl]propanoic acid (31)

^{1}H NMR (400 Mhz, DMSO) δ 12.17 (br s, 1H), 11.74 (s, 1H), 8.52 (d, J = 8.9 Hz, 1H), 7.59 – 7.51 (m, 2H), 7.47 – 7.39 (m, 1H), 7.38 – 7.32 (m, 1H), 7.32 – 7.26 (m, 2H), 6.93 – 6.87 (m, 2H), 5.84 (s, 1H), 5.42 – 5.33 (m, 1H), 2.92 (dd, J = 15.8, 8.2 Hz, 1H), 2.76 (dd, J = 15.8, 6.2 Hz, 1H), 1.27 (s, 9H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{23}H_{24}FN_3O_5$: 441.1695; Found 441.1693 $[M+H]^+$.

(3R)-3-(4-fluorophenyl)-3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (32)

^{1}H NMR (400 Mhz, DMSO) δ 12.26 (br s, 1H), 11.97 (br s, 1H), 8.61 (d, J = 8.6 Hz, 1H), 7.49 – 7.33 (m, 5H), 7.19 – 7.10 (m, 2H), 6.96 – 6.91 (m, 1H), 5.85 (s, 1H), 5.44 – 5.35 (m, 1H), 3.81 (s, 3H), 2.96 (dd, J = 15.8, 8.1 Hz, 1H), 2.81 (dd, J = 15.8, 6.4 Hz, 1H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_5$: 399.1225; Found 399.1228 $[M+H]^+$.

3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-(3-methylphenyl)propanoic acid (33)

^{1}H NMR (400 Mhz, DMSO) δ 12.21 (br s, 1H), 11.74 (br s, 1H), 8.48 (d, J = 8.7 Hz, 1H), 7.67 – 7.59 (m, 2H), 7.23 – 7.16 (m, 3H), 7.08 – 7.01 (m, 3H), 5.84 (s, 1H), 5.41 – 5.32 (m, 1H), 3.80 (s, 3H), 2.94 (dd, J = 15.9, 8.1 Hz, 1H), 2.78 (dd, J = 15.9, 6.3 Hz, 1H), 2.28 (s, 3H).

^{13}C NMR (176 MHz, DMSO) δ 172.06, 160.68, 157.88, 153.10, 145.22, 142.57, 137.17, 131.29, 128.12, 127.54, 127.28, 123.98 (2 C), 123.71, 114.00 (2 C), 87.13, 55.37, 48.97, 40.07, 21.08.

HRMS (m/z): $[M+H]^+$ calc. for $C_{21}H_{21}N_3O_5$: 395.1476; Found 395.1476 $[M+H]^+$.

3-(4-chlorophenyl)-3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (34)

^{1}H NMR (400 Mhz, DMSO) δ 12.28 (br s, 1H), 11.74 (s, 1H), 8.59 (d, J = 8.8 Hz, 1H), 7.66 – 7.59 (m, 2H), 7.46 – 7.35 (m, 4H), 7.08 – 7.01 (m, 2H), 5.83 (s, 1H), 5.42 – 5.33 (m, 1H), 3.80 (s, 3H), 2.96 (dd, J = 16.0, 8.0 Hz, 1H), 2.81 (dd, J = 16.0, 6.2 Hz, 1H).

^{13}C NMR (176 MHz, DMSO) δ 171.90, 160.83, 157.90, 153.16, 145.09, 141.67, 131.46, 131.27, 128.62 (2 C), 128.14 (2 C), 124.03 (2 C), 113.99, 87.12, 55.37, 48.52, 40.00.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}ClN_3O_5$: 415.0930; Found 415.0926 $[M+H]^+$.

(3R)-3-(4-fluorophenyl)-3-[[1-(4-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (35)

^{1}H NMR (400 Mhz, DMSO) δ 12.26 (br s, 1H), 12.00 (br s, 1H), 8.61 (d, J = 8.8 Hz, 1H), 7.83 – 7.75 (m, 2H), 7.48 – 7.41 (m, 2H), 7.39 – 7.31 (m, 2H), 7.18 – 7.10 (m, 2H), 5.85 (s, 1H), 5.44 – 5.35 (m, 1H), 2.96 (dd, J = 15.9, 8.2 Hz, 1H), 2.80 (dd, J = 15.9, 6.3 Hz, 1H).

^{13}C NMR (176 MHz, DMSO) δ 171.93, 161.13 (d: J = 242.5 Hz), 160.64, 160.35 (d: J = 243.8 Hz), 153.60, 145.68, 138.82 (d: J = 2.9 Hz), 134.64 (d: J = 2.7 Hz), 128.66 (d: J = 8.1 Hz, 2 C), 124.21 (d: J = 8.5 Hz, 2 C), 115.71 (d: J = 22.8 Hz, 2 C), 114.90 (d: J = 21.2 Hz, 2 C), 87.26, 48.47, 40.06.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}F_2N_3O_4$: 387.1025; Found 387.1019 $[M+H]^+$.

3-[[1-(4-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-[4-[(2-methylpropan-2-yl)oxy]phenyl]propanoic acid (36)

^{1}H NMR (400 Mhz, DMSO) δ 12.18 (br s, 1H), 12.01 (br s, 1H), 8.55 (d, J = 8.9 Hz, 1H), 7.84 – 7.75 (m, 2H), 7.39 – 7.27 (m, 4H), 6.95 – 6.87 (m, 2H), 5.86 (s, 1H), 5.44 – 5.34 (m, 1H), 2.93 (dd, J = 15.9, 8.2 Hz, 1H), 2.79 (dd, J = 15.9, 6.2 Hz, 1H), 1.27 (s, 9H).

^{13}C NMR (125 MHz, DMSO) δ 172.08, 160.54, 160.34 (d: J = 243.7 Hz), 153.94, 153.57, 145.76, 137.05, 134.65 (d: J = 2.6 Hz), 127.27 (2 C), 124.17 (d: J = 8.5 Hz, 2 C), 123.19 (2C), 115.71 (d: J = 22.8 Hz, 2 C), 87.29, 77.74, 48.43, 40.12, 28.52 (3 C).

HRMS (m/z): $[M+H]^+$ calc. for $C_{23}H_{24}FN_3O_5$: 441.1695; Found 441.1689 $[M+H]^+$.

3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-[4-[(2-methylpropan-2-yl)oxy]phenyl]propanoic acid (37)

^{1}H NMR (400 Mhz, DMSO) δ 12.25 (br s, 1H), 11.97 (br s, 1H), 8.55 (d, J = 8.9 Hz, 1H), 7.44 – 7.27 (m, 5H), 6.96 – 6.88 (m, 3H), 5.86 (s, 1H), 5.44 – 5.35 (m, 1H), 3.81 (s, 3H), 2.94 (dd, J = 15.9, 8.3 Hz, 1H), 2.80 (dd, J = 15.9, 6.2 Hz, 1H), 1.27 (s, 9H).

Compound only has a low purity,13 C data therefore not included. Compound is a 6.4:1 of tBu and OH phenol.

HRMS (m/z): $[M+H]^+$ calc. for $C_{24}H_{27}N_3O_6$: 453.1894; Found 453.1904 $[M+H]^+$.

3-(4-tert-butylphenyl)-3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (38)

^{1}H NMR (400 Mhz, DMSO) δ 12.60 – 11.60 (m, 2H), 8.54 (d, J = 8.9 Hz, 1H), 7.45 – 7.29 (m, 7H), 6.96 – 6.90 (m, 1H), 5.85 (s, 1H), 5.43 – 5.34 (m, 1H), 3.81 (s, 3H), 2.95 (dd, J = 15.9, 8.1 Hz, 1H), 2.81 (dd, J = 15.9, 6.3 Hz, 1H), 1.25 (s, 9H).

^{13}C NMR (176 MHz, DMSO) δ 172.12, 160.58, 159.52, 153.60, 149.21, 145.66, 139.55, 139.30, 129.78, 126.33 (2 C), 124.95 (2 C), 114.12, 111.78, 108.09, 87.45, 55.34, 48.66, 39.95, 34.11, 31.10 (3 C).

HRMS (m/z): $[M+H]^+$ calc. for $C_{24}H_{27}N_3O_5$: 437.1945; Found 437.1941 $[M+H]^+$.

(3R)-3-(4-fluorophenyl)-3-[[1-(3-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (39)

^{1}H NMR (400 Mhz, DMSO) δ 12.27 (br s, 1H), 8.66 (d, J = 8.8 Hz, 1H), 7.75 – 7.66 (m, 2H), 7.58 – 7.51 (m, 1H), 7.48 – 7.42 (m, 2H), 7.22 – 7.11 (m, 3H), 5.86 (s, 1H), 5.44 – 5.36 (m, 1H), 2.96 (dd, J = 15.9, 8.3 Hz, 1H), 2.82 (dd, J = 15.9, 6.3 Hz, 1H). COOH not observed due to fast exchange with water.

^{13}C NMR (125 MHz, DMSO) δ 171.92, 162.05 (d: J = 243.0 Hz), 161.14 (d: J = 242.5 Hz), 160.51, 154.09, 146.01, 139.78 (d: J = 10.7 Hz), 138.78 (d: J = 3.0 Hz), 130.77 (d: J = 9.2 Hz), 128.65 (d: J = 8.1 Hz, 2 C), 117.17 (d: J = 2.8 Hz), 114.92 (d: J = 21.2 Hz, 2 C), 113.07 (d: J = 21.0 Hz), 108.53 (d: J = 26.2 Hz), 87.57, 48.49, 40.08.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}F_2N_3O_4$: 387.1025; Found 387.1022 $[M+H]^+$.

(3R)-3-[[1-(3-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-(4-methylphenyl)propanoic acid (40)

^{1}H NMR (400 Mhz, DMSO) δ 12.25 (br s, 1H), 12.23 (br s, 1H), 8.57 (d, J = 8.8 Hz, 1H), 7.75 – 7.67 (m, 2H), 7.58 – 7.50 (m, 1H), 7.31 – 7.26 (m, 2H), 7.22 – 7.15 (m, 1H), 7.14 – 7.10 (m, 2H), 5.86 (s, 1H), 5.41 – 5.33 (m, 1H), 2.94 (dd, J = 15.8, 8.2 Hz, 1H), 2.79 (dd, J = 15.8, 6.4 Hz, 1H), 2.26 (s, 3H).

^{13}C NMR (125 MHz, DMSO) δ 172.05, 162.06 (d: J = 243.0 Hz), 160.43, 154.12, 146.11, 139.83 (d: J = 10.7 Hz), 139.55, 136.00, 130.77 (d: J = 9.2 Hz), 128.74 (2 C), 126.54 (2 C), 117.10, (d: 2.7 Hz), 113.01 (d: 22.9 Hz), 108.47 (d: J = 26.2 Hz), 87.54, 48.78, 40.12, 20.60.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_4$: 383.1276; Found 383.127 $[M+H]^+$.

(3R)-3-[[1-(4-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-(4-methylphenyl)propanoic acid (41)

^{1}H NMR (400 Mhz, DMSO) δ 12.21 (br s, 1H), 12.00 (s, 1H), 8.52 (d, J = 8.8 Hz, 1H), 7.83 – 7.76 (m, 2H), 7.38 – 7.25 (m, 4H), 7.14 – 7.09 (m, 2H), 5.85 (s, 1H), 5.41 – 5.32 (m, 1H), 2.93 (dd, J = 15.9, 8.0 Hz, 1H), 2.78 (dd, J = 15.9, 6.3 Hz, 1H), 2.26 (s, 3H).

^{13}C NMR (125 MHz, DMSO) δ 172.06, 160.56, 160.31 (d: J = 243.7 Hz), 153.62, 145.77, 139.57, 135.98, 134.68 (d: 2.7), 128.73 (2 C), 126.55 (2 C), 124.12 (d: J = 8.5 Hz, 2 C), 115.70 (d: J = 22.8 Hz, 2 C), 87.2, 48.76, ~40.1 (below solvent signal), 20.60.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_4$: 383.1276; Found 383.1274 $[M+H]^+$.

(3R)-3-[[5-hydroxy-1-(3-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-(4-methylphenyl)propanoic acid (42)

^{1}H NMR (400 Mhz, DMSO) δ 12.50 – 11.70 (m, 2H), 8.52 (d, J = 9.0 Hz, 1H), 7.44 – 7.33 (m, 3H), 7.31 – 7.25 (m, 2H), 7.15 – 7.09 (m, 2H), 6.96 – 6.90 (m, 1H), 5.85 (s, 1H), 5.42 – 5.32 (m, 1H), 3.80 (s, 3H), 2.93 (dd, J = 15.8, 8.1 Hz, 1H), 2.79 (dd, J = 15.8, 6.2 Hz, 1H), 2.26 (s, 3H).

^{13}C NMR (125 MHz, DMSO) δ 172.10, 160.57, 159.53, 153.62, 145.66, 139.56, 139.30, 135.99, 129.79, 128.74 (2 C), 126.55 (2 C), 114.12, 111.80, 108.08, 87.47, 55.34, 48.75, 40.08, 20.60

HRMS (m/z): $[M+H]^+$ calc. for $C_{21}H_{21}N_3O_5$: 395.1476; Found 395.1480 $[M+H]^+$.

3-(4-tert-butylphenyl)-3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (43)

^{1}H NMR (400 Mhz, DMSO) δ 12.21 (s, 1H), 11.73 (s, 1H), 8.48 (d, J = 8.9 Hz, 1H), 7.65 – 7.59 (m, 2H), 7.35 – 7.30 (m, 4H), 7.07 – 7.01 (m, 2H), 5.83 (s, 1H), 5.42 – 5.33 (m, 1H), 3.80 (s, 3H), 2.94 (dd, J = 15.8, 8.0 Hz, 1H), 2.80 (dd, J = 15.8, 6.3 Hz, 1H), 1.25 (s, 9H).

^{13}C NMR (125 MHz, DMSO) δ 172.11, 160.69, 157.87, 153.12, 149.18, 145.24, 139.60, 131.31, 126.35 (2 C), 124.94 (2 C), 123.96 (2 C), 113.99 (2 C), 87.10, 55.37, 48.64, ~39.9 (below solvent signal), 34.12, 31.11.

HRMS (m/z): $[M+H]^+$ calc. for $C_{24}H_{27}N_3O_5$: 437.1945; Found 437.1936 $[M+H]^+$.

3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]-3-[4-[(2-methylpropan-2-yl)oxy]phenyl]propanoic acid (44)

^{1}H NMR (400 Mhz, DMSO) δ 12.16 (br s, 1H), 11.74 (br s, 1H), 8.50 (d, J = 8.9 Hz, 1H), 7.65 – 7.60 (m, 2H), 7.33 – 7.28 (m, 2H), 7.07 – 7.02 (m, 2H), 6.93 – 6.88 (m, 2H), 5.84 (s, 1H), 5.43 – 5.34 (m, 1H), 3.80 (s, 3H), 2.93 (dd, J = 15.8, 8.1 Hz, 1H), 2.79 (dd, J = 15.8, 6.2 Hz, 1H), 1.27 (s, 9H).

Compound only has a low purity,13 C data therefore not included. Compounds is a 6.7:1 mixture of tBu and OH Phenol.

HRMS (m/z): $[M+H]^+$ calc. for $C_{24}H_{27}N_3O_6$: 453.1894; Found 453.1905 $[M+H]^+$.

(3R)-3-(4-fluorophenyl)-3-[[1-(2-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]propanoic acid (45)

1H NMR (400 Mhz, DMSO) δ 12.22 (br s, 1H), 11.75 (br s, 1H), 8.59 (d, J = 8.7 Hz, 1H), 7.59 – 7.51 (m, 2H), 7.47 – 7.40 (m, 3H), 7.38 – 7.32 (m, 1H), 7.17 – 7.09 (m, 2H), 5.83 (s, 1H), 5.42 – 5.34 (m, 1H), 2.94 (dd, J = 16.0, 8.2 Hz, 1H), 2.78 (dd, J = 16.0, 6.5 Hz, 1H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{19}H_{15}F_2N_3O_4$: 387.1025; Found 387.1021 $[M+H]^+$.

(3R)-3-[[1-(2-fluorophenyl)-5-hydroxypyrazole-3-carbonyl]amino]-3-(4-methylphenyl)propanoic acid (46)

^1H NMR (400 Mhz, DMSO) δ 12.16 (br s, 1H), 11.73 (br s, 1H), 8.49 (d, J = 8.8 Hz, 1H), 7.59 – 7.51 (m, 2H), 7.47 – 7.40 (m, 1H), 7.38 – 7.32 (m, 1H), 7.29 – 7.24 (m, 2H), 7.13 – 7.08 (m, 2H), 5.83 (s, 1H), 5.39 – 5.31 (m, 1H), 2.91 (dd, J = 15.9, 8.1 Hz, 1H), 2.75 (dd, J = 15.9, 6.4 Hz, 1H), 2.25 (s, 3H).

The amount was insufficient for ^{13}C NMR analysis.

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_4$: 383.1276; Found 383.1265 $[M+H]^+$.

(3R)-3-(4-fluorophenyl)-3-[[5-hydroxy-1-(4-methoxyphenyl)pyrazole-3-carbonyl]amino]propanoic acid (47)

^1H NMR (400 Mhz, DMSO) δ 12.26 (br s, 1H), 11.74 (br s, 1H), 8.57 (d, J = 8.7 Hz, 1H), 7.65 – 7.59 (m, 2H), 7.48 – 7.41 (m, 2H), 7.18 – 7.10 (m, 2H), 7.08 – 7.01 (m, 2H), 5.83 (s, 1H), 5.43 – 5.35 (m, 1H), 3.80 (s, 3H), 2.96 (dd, J = 15.9, 8.2 Hz, 1H), 2.80 (dd, J = 15.9, 6.4 Hz, 1H).

^{13}C NMR (125 MHz, DMSO) δ 171.96, 161.12 (d: J = 242.5 Hz), 160.77, 157.91, 153.13, 145.15, 138.86 (d: J = 3.0 Hz), 131.27, 128.67 (d: J = 8.1 Hz, 2 C), 124.04 (2 C), 114.88 (d: J = 21.2 Hz, 2C), 114.00 (2 C), 87.14, 55.38, 48.44, ~40.0 (below solvent signal).

HRMS (m/z): $[M+H]^+$ calc. for $C_{20}H_{18}FN_3O_5$: 399.1225; Found 399.1225 $[M+H]^+$.

Prompts used for RIPK1 studies

Prompt2:

You are an expert in computational and medicinal chemistry. You can generate valid SMILES for novel and structurally diverse molecules with the desired properties.

From the following examples learn the structural patterns that could lead to high experimental activity.

Examples:

SMILES, Experimental activity, Molecular weight (MW):

{examples}

Rules to apply:

1. The generated molecule contains the core 'O=C1NC=NN1' or 'O=C1NC=CN1'

2. The core should have at least two substituents

3. Molecular weight of the generated molecule is between 350 and 450 Daltons

4. The generated molecule is not part of the examples

Task: Using the learned patterns and strictly obeying all the rules, modify the following SMILES to generate {nmol} novel and structurally similar valid SMILES for molecules with similar experimental activity. Please provide a confidence score that goes from 0 to 10 depending on how confident you feel about the predicted activity.

Output the generated SMILES string, their predicted activity and molecular weight (MW), and confidence score in TSV format with 'SMILES', 'predicted_activity', 'predicted_MW', and 'Confidence' columns.

Input SMILES: {input_smiles}

Output format: table with 'SMILES' and 'predicted_activity' columns in TSV format. No explanation. Only the table is required.

Prompt3-keto:

You are an expert in computational and medicinal chemistry. You can generate valid SMILES for novel and structurally diverse molecules with the desired properties.

From the following examples learn the structural patterns that could lead to high experimental activity.

Examples:

SMILES, Experimental activity, Molecular weight (MW):

{examples}

Rules to apply:

1. The generated molecule contains the core 'N1C=CNC1=O' or 'N1C=NNC1=O'

2. The core is the central part of the molecule and should have at least two N-substituents

3. The generated molecule can contain a piperidine

4. Molecular weight of the generated molecule is between 380 and 480 Daltons

5. The generated molecule is not part of the examples

Task: Using the learned patterns and strictly obeying all the rules, modify the following SMILES to generate {nmol} novel and structurally similar valid SMILES for molecules with similar experimental activity. Please provide a confidence score that goes from 0 to 10 depending on how confident you feel about the predicted activity.

Output the generated SMILES string, their predicted activity and molecular weight (MW), and confidence score in TSV format with 'SMILES', 'predicted_activity', 'predicted_MW', and 'Confidence' columns.

Input SMILES: {input_smiles}

Output format: table with 'SMILES' and 'predicted_activity' columns in TSV format. No explanation. Only the table is required.

Prompt3-pyro:

You are an expert in computational and medicinal chemistry. You can generate valid SMILES for novel and structurally diverse molecules with the desired properties.

From the following examples learn the structural patterns that could lead to high experimental activity.

Examples:

SMILES, Experimental activity, Molecular weight (MW):

{examples}

Rules to apply:

1. The generated molecule contains the core 'C1CCCN1C=O' or 'C1CCN1C=O'

2. The core is the central part of the molecule and should have at least two substituents, with at least one aromatic ring

3. The generated molecule can contain a piperidine

4. Molecular weight of the generated molecule is between 380 and 480 Daltons

5. The generated molecule is not part of the examples

Task: Using the learned patterns and strictly obeying all the rules, modify the following SMILES to generate {nmol} novel and structurally similar valid SMILES for molecules with similar experimental activity. Please provide a confidence score that goes from 0 to 10 depending on how confident you feel about the predicted activity.

Output the generated SMILES string, their predicted activity and molecular weight (MW), and confidence score in TSV format with 'SMILES', 'predicted_activity', 'predicted_MW', and 'Confidence' columns.

Input SMILES: {input_smiles}

Output format: table with 'SMILES' and 'predicted_activity' columns in TSV format. No explanation. Only the table is required.

Prompt4-keto:

You are an expert in computational and medicinal chemistry. You can generate valid SMILES for novel and structurally diverse molecules with the desired properties.

From the following examples learn the structural patterns that could lead to high experimental activity.

Examples:

SMILES, Experimental activity, Molecular weight (MW):

{examples}

Rules to apply:

1. The generated molecule contains the core 'N1C=CNC1=O' or 'N1C=NNC1=O'

2. The core is the central part of the molecule and should have at least two N-substituents

3. The generated molecule should contain a piperidine

4. Molecular weight of the generated molecule is between 380 and 480 Daltons

5. The generated molecule is not part of the examples

Task: Using the learned patterns and strictly obeying all the rules, modify the following SMILES to generate {nmol} novel and structurally similar valid SMILES for molecules with similar experimental activity. Please provide a confidence score that goes from 0 to 10 depending on how confident you feel about the predicted activity.

Output the generated SMILES string, their predicted activity and molecular weight (MW), and confidence score in TSV format with 'SMILES', 'predicted_activity', 'predicted_MW', and 'Confidence' columns.

Input SMILES: {input_smiles}

Output format: table with 'SMILES' and 'predicted_activity' columns in TSV format. No explanation. Only the table is required.

Prompt4-pyro:

You are an expert in computational and medicinal chemistry. You can generate valid SMILES for novel and structurally diverse molecules with the desired properties.

From the following examples learn the structural patterns that could lead to high experimental activity.

Examples:

SMILES, Experimental activity, Molecular weight (MW):

{examples}

Rules to apply:

1. The generated molecule contains the core 'C1CCCN1C=O' or 'C1CCN1C=O'

2. The core is the central part of the molecule and should have at least two substituents, with at least one aromatic ring

3. The generated molecule should contain a piperidine

4. Molecular weight of the generated molecule is between 380 and 480 Daltons

5. The generated molecule is not part of the examples

Task: Using the learned patterns and strictly obeying all the rules, modify the following SMILES to generate {nmol} novel and structurally similar valid SMILES for molecules with similar experimental activity. Please provide a confidence score that goes from 0 to 10 depending on how confident you feel about the predicted activity.

Output the generated SMILES string, their predicted activity and molecular weight (MW), and confidence score in TSV format with 'SMILES', 'predicted_activity', 'predicted_MW', and 'Confidence' columns.

Input SMILES: {input_smiles}

Output format: table with 'SMILES' and 'predicted_activity' columns in TSV format. No explanation. Only the table is required.

Prompts used for CatA studies

Prompt:

You are an expert in computational and medicinal chemistry. You can generate valid SMILES for novel and structurally diverse molecules with the desired properties.

From the following examples learn the structural patterns that could lead to high experimental activity.

Examples:

SMILES, Experimental activity, Molecular weight (MW):

{examples}

Rules to apply:

1. The generated molecule contains pyrazolone core (SMILES: 'C1=CN=NC1=O')

2. Molecular weight of the generated molecule is lower than 430 Daltons

3. Maximal number of chemical bonds to any aromatic ring is two, the only exception can be the pyrazolone core

4. Only chemical species with a single fluorine are allowed, for instance there are no difluoro- and trifluoro- moieties and no '(F)F' and '(F)(F)' patterns in the SMILES

5. There is no iodine and bromine in the generated molecule

Task: Using the learned patterns and strictly obeying all the rules, modify the following SMILES to generate {nmol} novel and structurally similar valid SMILES for molecules with higher experimental activity. Please provide a confidence score that goes from 0 to 10 depending on how confident you feel about the predicted activity.

Output the generated SMILES string, their predicted activity and molecular weight (MW), and confidence score in TSV format with 'SMILES', 'predicted_activity', 'predicted_MW', and 'Confidence' columns.

Input SMILES: {input_smiles}

Output format: table with 'SMILES' and 'predicted_activity' columns in TSV format. No explanation. Only the table is required.