Harnessing Medicinal Chemical Intuition from Collective Intelligence

Pierre Llompart[1, 2, *], Kwame Amaning[1], Marc Bianciotto[1], Bruno Filoche-Rommé[1], Yann Foricher[1], Pablo Mas[1,3], David Papin[1], Jean-Philippe Rameau[1], Laurent Schio[1], Gilles Marcou[2], Alexandre Varnek[2], Mehdi Moussaid[4,5], Claire Minoletti[1, *], Paraskevi Gkeka[1, *]

[1]Integrated Drug Discovery, Sanofi, Vitry-sur-Seine, France

[2]Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg, France

[3] Theoretical Chemistry Department, École Normale Supérieure, Paris, France

[4]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[5]School of Collective Intelligence, Université Mohammed VI Polytechnique, Rabat, Morocco

*Corresponding authors: Paraskevi.Gkeka@sanofi.com, mailto:Pierre.Llompart@sanofi.com, Claire.Minoletti@sanofi.com

Keywords: medicinal chemistry intuition, collective intelligence, ADMET, drug design, GNN, lead optimization

Abstract

Over the last decade, the combination of collective intelligence with computational methods has transformed complex problem-solving. Here, we investigate if and how collective intelligence can be applied to drug discovery, focusing on the lead optimization stage of the discovery process. For this study, 92 Sanofi researchers with diverse scientific expertise participated anonymously in a lead optimization exercise. Their feedback was used to build a collective intelligence agent that was compared to an artificial intelligence model developed in parallel. This work has led to three major conclusions. First, a significant improvement of collective versus individual decisions in optimizing ADMET endpoints is observed. Second, for all endpoints apart from hERG inhibition, the collective intelligence performance exceeds the artificial intelligence model. Third, we observe a complementarity between collective intelligence and AI for complex tasks, demonstrating the potential of hybrid predictions. Overall, this research highlights the potential of collective intelligence in drug discovery. The entire dataset, including questionnaire responses, and developed models are available for access on GitHub.

Introduction

Chemical intuition can be defined as the ability of experienced chemists to anticipate the outcomes of chemical reactions, predict molecular interactions, and envisage the impact of structural modifications on a compound's properties. This intuition, honed through years of practice, guides chemists in the complex, multi-step process of drug discovery. During the ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) drug optimization stage, medicinal chemistry intuition is often employed to estimate the pharmacokinetic properties of a molecule based on its similarity to known compounds. In an industrial setting, this intuition combined with *in silico* property prediction models drive the multiparametric lead optimization process.[1,2] Recently, the responses of 35 chemists on binary medicinal chemistry questions were provided as input to an artificial intelligence (AI) learning-to-rank framework. This work led to the development of an implicit drug-likeliness scoring function, able to capture aspects of chemistry not covered by other computational counterparts, *i.e.*, metrics and rule sets.[3]

While medicinal chemistry intuition in drug discovery relies heavily on individual experience, know-how and personal bias,[4,5,6] collective intelligence (CI), *i.e.*, the capacity of a group to solve complex problems, has shown considerable improvement in reinforcing human decision-making.[7] Collective intelligence thrives on significant group size, participants diversity as well as different data aggregation methods. It can outperform the capacities of individual group members and even surpass experts in complex tasks.[7,8] CI effectiveness lies in its ability to merge multiple viewpoints into a cohesive answer, thereby mitigating the impact of individual decision biases, reducing noise, and harnessing the plurality of ideas, knowledge bases, and cognitive approaches.

Collective intelligence and chemical intuition have already been combined in the fields of metal-organic frameworks[9] and inorganic chemistry experiments.[10] Nevertheless, the most striking scientific results have been produced for the prediction of biological structures through the Foldit initiative.[11] As of now, Foldit has been applied to other related fields such as small molecule and protein design.[12] Recently, a similar crowdsourcing approach was adopted for RNA design and folding prediction.[13,14] Only a few examples exist where collective intelligence has been applied to the decision making in drug discovery.[3,15] Often drug discovery and in particular the stage of lead optimization relies heavily on singular experts or small project teams, however, as illustrated by Hong and Page,[8] groups with diverse perspectives can outperform like-minded experts. In the context of drug design, this diversity, referred to as heterogenous collective intelligence, could yield a more efficient process.

To deliver a unified and reliable approach for compound prioritization, we investigated whether the application of collective intelligence can improve the decision-making process in ADMET optimization. We conducted an experiment with Sanofi scientists with an expertise ranging from Pharmacokinetics, Structural and In vitro Biology to Molecular Modeling and Medicinal Chemistry. This endeavor aimed to compare the performance of collective versus individual input. By employing various aggregation methods, we aimed to understand the critical factors influencing collective intelligence' success rate (SR), such as the confidence of the participant in a specific answer and their medicinal chemistry expertise. Furthermore, we examined potential medicinal chemistry pitfalls from collective decisions that might lead to optimization bias. We finally sought to assess the performance and potential support from an AI model to individual and collective decision-making processes.

Results

This section is organized as follows. First, we provide an overview of the data we obtained from our experiment, and we illustrate the existence of a correlation between the responses of the participants and their medicinal chemistry background and confidence per response. Second, we assess the performance of collective intelligence per ADMET endpoint, examining how team composition and aggregation method influence the collective outcomes. Finally, we explore the collective biases encountered in medicinal chemistry during the exercise and underscore the complementarity between AI and CI responses.

Overview of the experiment and analysis of the collective intelligence

During the experiment, 92 participants with diverse scientific background and roles in drug discovery were asked 74 ADMET optimization questions (**Figure 1a**). Participants self-rated their medicinal chemistry expertise on a scale from level 1, corresponding to minimal knowledge, to level 5, corresponding to medicinal chemistry experts (**Figure 1b**). Due to technical limitations, the experiment was divided into two sessions of 37 questions each. For each question participants were given a chemical scaffold and asked to choose the 'best' of three proposed substituents for a specified ADMET endpoint (**Figure S1**). Additionally, for every answer, participants were required to rate their confidence in the selected substituent from 1 (low) to 5 (high). The experiment yielded a total of 6,808 responses and their corresponding confidence levels.

The median of the global performance defined as success rate (SR), *i.e.*, correct responses over the total number of questions, was 43%, while we observed a lowest and highest success rate of 8% and 73%, respectively (median and outliers of blue violin plot in **Figure 1c**). Here, skill groups correspond to the expertise scale defined above and provided by each participant. Global median aligns closely with group 3 (43%) and is lower than groups 4 (52%) and 5 (58%). One out of four participants had a success rate of more than 50%. As a reminder, each question had three possible answers and the random success rate is expected to be 33%.

The Collective Intelligence (CI) response is defined as the answer selected using a "democratic" approach, *i.e.*, most frequent response per question. Globally, the CI response exceeded the median SR for all expertise-based groups as well as for the global group, by up to 18% (**Figure 1c**). Based on these SR, groups 1 and 2 and groups 4 and 5 can be merged into *'non-Experts'* and *'Experts'* cohorts, respectively. Nevertheless, participants from group 3 exhibit diverse performances, ranging from as low as 20% to 62% SR, demonstrating an unreliable self-evaluation in agreement with previous studies.[16] Interestingly, characterizing group 3 participants as '*Experts*' does not significantly affect the *'Expert'* SR which is 56±6% when only groups 4 and 5 are considered versus 52±7% when group 3 is also included (**Figures S2a and S2b**). To validate our choice to include group 3 in the 'Experts', we performed an *a posteriori* analysis where we classified as '*non-Experts'* all participants with individual SR less than 50% and those with SR more than 50% as 'Experts' (**Figure S2c**).

Individuals with higher self-assessed expertise displayed greater confidence in their answers (**Figure 1d and S3a**). For instance, 81% of level 5 experts assigned a confidence level greater than or equal to 3 (**Figure S3a**). In contrast, non-experts predominantly chose the lowest confidence value (**Figure S3a**), and no correlation was shown between their SR and confidence (**Figure 1d**). However, a confidence level above 3 combined with an expertise level above 2 consistently led to a SR exceeding 50%, highlighting the significant effect of confidence and expertise combined on achieving high success rates.

The questions chosen for this study focused on five endpoints: the partition coefficient (logP), distribution coefficient (logD), aqueous solubility (logS), apparent permeability (P_{app}), and hERG inhibition (**Figure 1e**). Over half of the questions are related to aqueous solubility and distribution coefficient. Significant variations in the SR are observed for the different endpoints. While logP, permeability, and solubility endpoints achieved higher median SR of ~40% or more, hERG and logD present median SR close to the random benchmark of 33% (**Figure 1f and S4**). CI demonstrated effectiveness across most of these endpoints. For logP, it remarkably exceeded the median individual SR, achieving 100% with all participants. Similarly, for solubility and permeability, CI improved SR by ~20% in both cases (white-filled circles, **Figure 1f**). However, for hERG CI did not enhance performance, while for logD and SR improvement of approximately 10% was observed.

Interestingly, the prevalence of low and medium confidence responses was uniform across different endpoints (**Figures 1g and S3b**). This suggests that confidence proportions are more influenced by the characteristics of the group rather than the specific endpoint. Nonetheless, for endpoints like logP and permeability, there is a distinct correlation between confidence levels and SR. Conversely, for the hERG endpoint, no such correlation was apparent, indicating that in more complex problems, confidence may not be the key determinant of high SR.

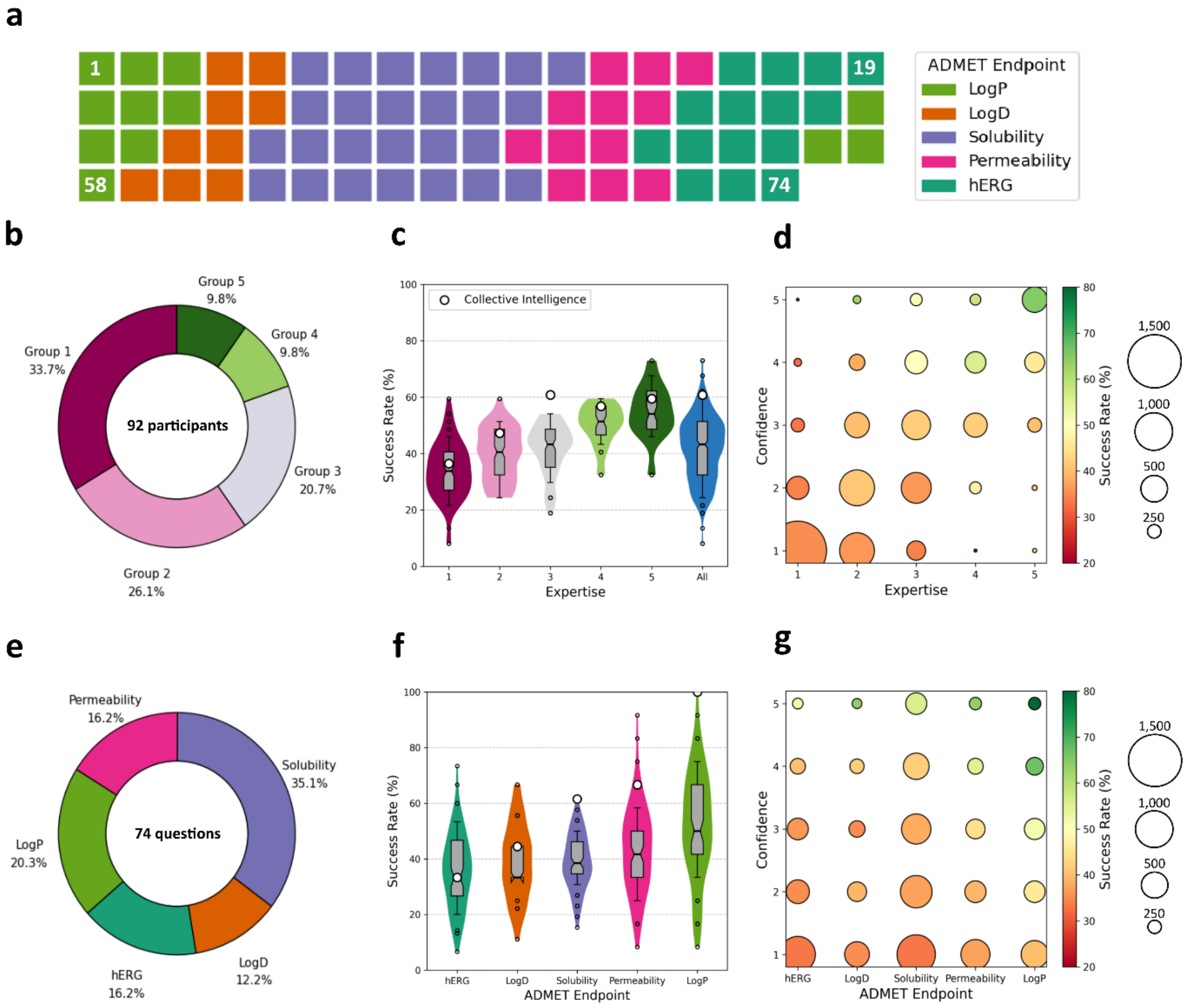


Figure 1: *Overview of the collective intelligence experiment and main results.* **a)** Distribution of the questions in the questionnaire. Each box corresponds to a question colored by endpoint. **b)** Participants' partitioning per self-labeled medicinal chemistry expertise. **c)** Violin plots of the SR by expertise level for each group (color code as in 1b) and all participants (blue). The median is shown as a horizontal line across the thinnest part of the boxes. The error bars correspond to the interquartile range. The collective SR are shown as white-filled circles, while the outliers are depicted as small circles. **d)** Bubble plot of the number of responses (size) and mean SR (color) dependence on medicinal chemistry self-labeling and confidence level per question. **e)** Endpoint distribution of the 74 questions. **f)** Violin plots of the SR

for each endpoint (colors as in e). The remaining plot features are as for c. **g)** Bubble plot of the number of responses (size) and mean SR (color) dependence on endpoint and confidence level per question.

To further investigate into these observations, we built a 2D map using the UMAP[17] method to visualize the participants' space based on their answers and levels of confidence (**Figure S5**). Each dot on the map represents a participant, with its position determined by the similarity in the participants' responses and confidence levels. Interestingly, experts and non-experts, occupy distinct areas with different SR. Nevertheless, participants self-labeled as level 2 or 3 are dispersed, and sometimes found in areas occupied by experts (**Figure S5a and S5b**). These results suggest significant noise in the self-assessed expertise levels. In the second part of the exercise (session two), the participants seem to have understood the difficulty of the questions and adjusted appropriately their level of expertise. This is indicated by the improved separation of expertise levels (**Figures S5c and S5d**).

Overall, our analyses demonstrate that the primary determinant correlating with the SR is the level of confidence to each question.

Collective performance dynamics and the effect of aggregation methods

This section of the study is focused on understanding how different aggregation methods affect the SR of CI in drug design, particularly considering factors such as confidence, expertise, and participants' population. To this end, we varied the sample size from a single individual (which in fact corresponds to the median SR of the participants) to the maximum group size of 92 participants. For each iteration, all unique combinations of individuals without permutations were analyzed to determine a distinct collective SR distribution. The evolution of this SR was analyzed based on the ADMET endpoint, expertise group, or aggregation method (see **Figure 2** and SI **Figures S6-S12**).

The collective responses were obtained using six different aggregation methods: the *'democratic'* approach (most frequent response), log confidence weighting (log odds), fuzzy logic aggregation,

https://doi.org/10.26434/chemrxiv-2024-0hww3 **ORCID:** https://orcid.org/0000-0002-0752-3539 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

confidence weighting, expertise weighting, and co-weighting by expertise and confidence. Using the democratic approach, the SR increased by 15% when going from a single to 20 participants, reaching a maximum of 60% when all participants are considered (**Figure 2a**). Expertise weighting nullified the effect of collective intelligence, while log confidence weighting achieved a 5% increase with only 15 participants compared to the most frequent SR. This effect is noticeable for smaller teams, but it becomes less pronounced and diminishes as the group size increases. These results indicate log confidence-based aggregation could be an effective aggregation method, enabling high collective SRs with smaller teams.

The evolution of CI SR was also analyzed across expertise groups (**Figures 2b and S6**). The non-expert group requires over 40 participants to reach a SR of approximately 50%, whereas the expert and mixed groups surpass 55% SR with only 10 (experts) and 15 (mixed) participants, respectively. This suggests that an effective CI team for drug design should ideally include some experts and consist of a minimum of 15 members. Notably, the CI SR difference between an all-expert team and a mixed team was minimal, with the mixed team requiring ~5 more participants to achieve comparable results using the log odds aggregation method (**Figure S6**).

The collective performance dynamics were also analyzed for each endpoint independently (**Figures 2c, 2d and S8-S13**). Using mixed-expertise teams, an 80% SR can be achieved with just ten participants for logP (**Figure 2c**). This trend also held for permeability and solubility (**Figures S8 and S9**). However, for hERG, over 70 participants were needed to exceed a 50% SR. A noticeable result is that in particular for hERG and only for the groups that include experts the best performing aggregation method is the one that accounts for the expertise level (**Figure 2d, S10**, and **S12**). A less pronounced and more unclear expertise influence was noted for logD (**Figures S11** and **S12**).

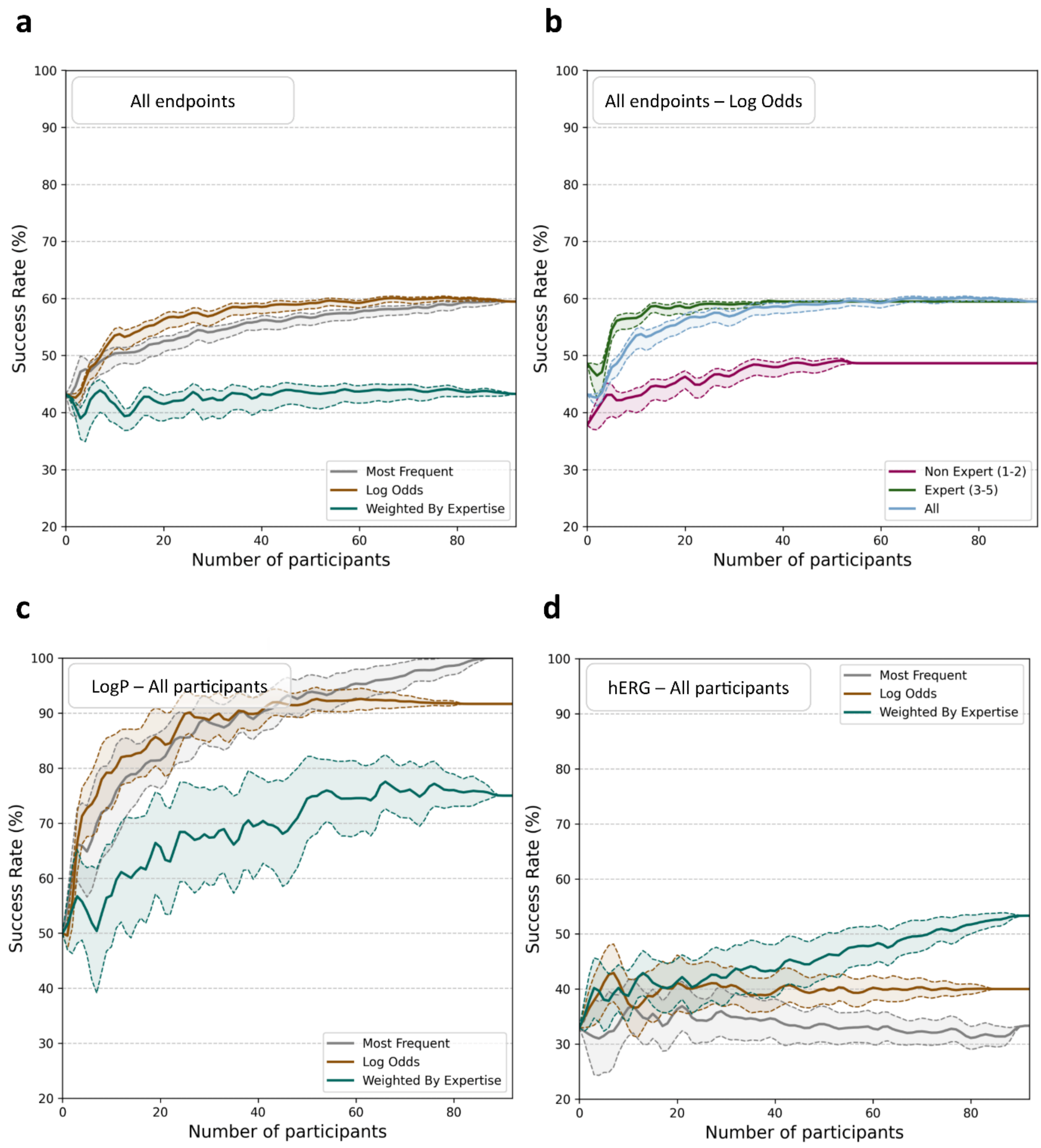


Figure 2: *Evolution of the collective success rate.* **a)** Collective SR for all endpoints using different aggregation methods**.** The collective answer is either obtained using the *'democratic'* (most frequent), the *confidence-weighted* (log odds), or the *expertise-weighted* aggregation method. **b)** Collective SR for a*ll endpoints* using l*og* odds for the different participants' groups. **c)** Collective SR for logP and for all participants. **d)** Collective SR for hERG and for all participants.

Medicinal chemistry pitfalls

This section aims to uncover potential biases in collective decision-making for optimization tasks. A 2D t-SNE[18] map was built to represent the chemical space, with each point corresponding to a compound (**Figure 3**). The set of all unique compounds used in the questionnaire is termed as *'CI library'*. This set, consisting of 193 compounds, was projected on the t-SNE map (**Figure 3**). The map reveals that no endpoint, among the five studied herein, occupies specific areas, indicating the diversity of the chemical set used despite its relatively small size. Some compounds overlap between endpoints as they were derived from the same research work, albeit they are not structurally identical. When examining the SR mapping, the performance appeared sparse, with only three distinct clusters of very low SR regrouping hERG and logD, or solubility and permeability (**Figure S13**).

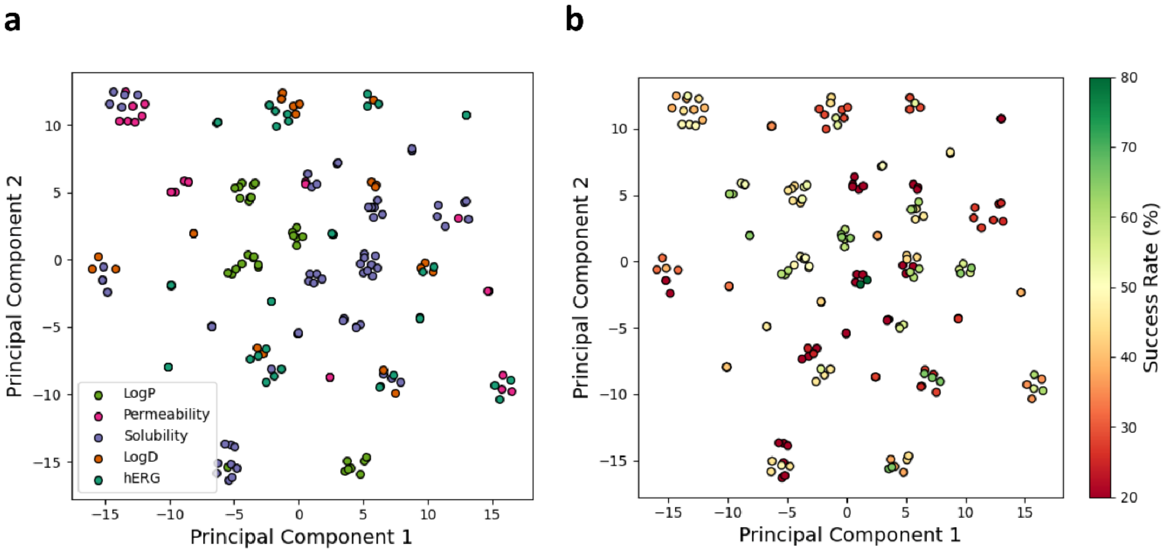
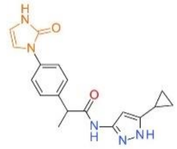
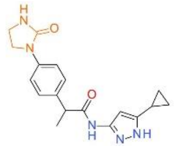
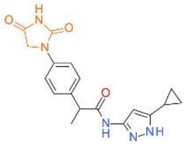
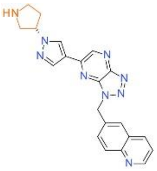
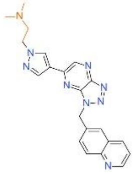
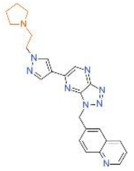
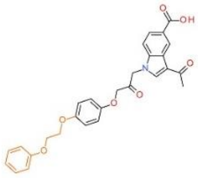
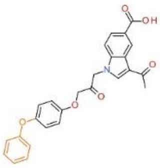
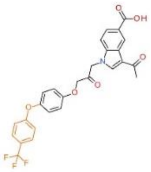
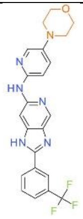
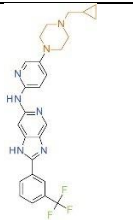
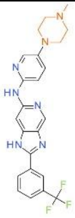
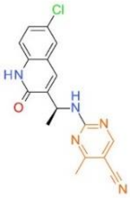
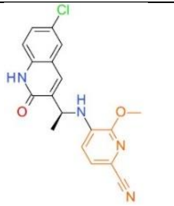
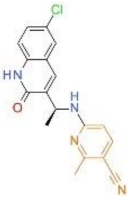
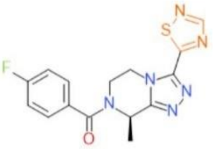
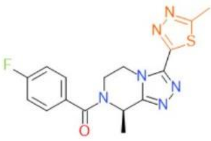
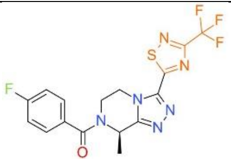
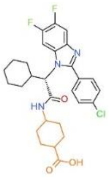
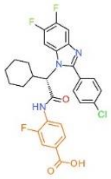
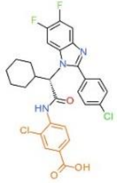
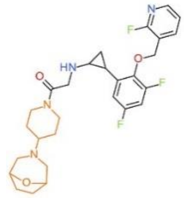
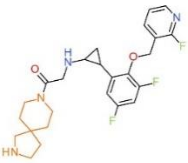
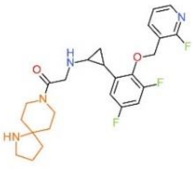


Figure 3: *t-SNE map of the CI library*. Each point represents a unique compound colored by **a)** ADMET endpoint and **b)** the success rate of the related question.

To better understand the collective errors or misconceptions of the participants, the worst-performing questions were examined. The structures selected by most participants were compared to the correct answers to try to investigate the origins of the errors (**Table 1**).

Table 1: *Selection of challenging cases from the CI questionnaire.* The table presents the collectively worst predicted examples, i.e., worst success rate per ADMET endpoint.

	ID	SR	Most given	Correct	Third Option	Ref.
Permeability	**1**	0.20				19
	2	0.26				20
Solubility	**3**	0.09				21
	4	0.12				22
LogD	**5**	0.19				23
	6	0.24				24
hERG	**7**	0.19				25
	8	0.26				26

There are multiple strategies to enhance permeability depending on whether it is a passive diffusion issue or linked to a transporter. Those strategies involve increasing lipophilicity and moving from ionizable groups to non-ionizable groups. Additional optimization plans are, for instance, reducing polarity, altering flexibility, replacing polar groups by isosters, lowering hydrogen bonding or favorizing intramolecular hydrogen bonding. In case **2** from **Table 1**, the correct answer was a tertiary amine, known to be less basic, but the majority of participants selected the pyrrolidine derivative. The observed difference versus the dimethylamine analogue was very difficult to expect, and *a posteriori* analysis did not permit to rationalize this result as the clogD of both derivatives were quite similar. For instance, in case **1**, it was challenging to foresee the impact of unsaturation to imidazolidinone on permeability.

For questions related to solubility improvement, the participants often favored polar compounds, adhering to the common knowledge that increase in polarity typically leads to better solubility.[27] Yet, in case **3**, the selected choice corresponds to compounds undergoing structural changes that impact dissolution more than polarity. Participants aimed at increasing the polarity linked to the ethylene glycol moiety, but the logP was similar between both compounds. Factors affecting solid state destabilization (*e.g.*, solid states, packing) might have weighted more and led to 3-4 fold difference in thermodynamic solubility. In case **4**, where substructures known for enhancing solubility competed, the corrected answer was the less common cyclopropyl-substituted N-methylene piperazine.[28] Responses were even split, when experts chose N-methyl piperazine derivative whereas naïve participants favored the morpholine analog.[28,29] None of the groups selected the corrected answer. In hindsight, medicinal chemists concurred that cyclopropyl can be considered as a bioisoster of alkene moiety or also a phenyl ring, and as such, it did not occur to be a strategy to improve solubility.[30] Notably, while logP rankings were accurately predicted, the misjudgment stemmed from evaluating with errors the weight of contributing phenomena, an effect exaggerated probably by stressful testing conditions.

Despite its direct relation to logP where CI was shown to be effective, logD-related collective decisions showed limitations. This effect arises from pKa's influence on logD, making it hard to predict a compound's ionization state, especially with multiple ionizable sites and mesomere interdependence.[31] Cases **5** and **6** exemplify this complexity, where altering nitrogen positions in the aromatic cycle and introducing groups like methyl or O-methyl complicates the intuitive evaluation of inductive effects beyond common rules.[32]

hERG affinity prediction also proved challenging, due to its dependence to both a compound's intrinsic properties and its interaction with the hERG channel, known for its flexibility and the structural diversity of its binders.[33] In case **7**, participants chose a compound with an acidic moiety on a saturated ring, a choice that had the double effect of favoring a more acidic moiety and likely avoiding π-π interactions typical in the hERG binding site. Nevertheless, the correct compound relied more on electronic effects on its aromatic moiety. For case **8**, the choice was guided by the characteristic of the molecule which displayed reduced basicity and steric hindrance with bridge moiety potentially leading to lower hERG affinity. Yet, the correct answer, demonstrating lower hERG affinity, corresponds to a molecule that likely adopts a non-traditional binding mode, diverging from expected interaction patterns.

Collective intelligence versus AI models

We also evaluated the performance of predictive models, specifically the ChemProp[34] Graph Neural Networks (GNNs), in similar decision tasks. GNN models, trained on curated public data (**Table S1, Figures S14** and **S15**), were used to respond to the CI questionnaire. Their objective was to predict endpoint measurements and select options leading to optimal values, *i.e.*, lower logP and logD, higher logS, permeability, and hERG pIC50. We first compared the GNNs' SR to both individual and confidence-weighted CI performances (**Figure 4a**). Subsequently, we assessed the potential of AI to enhance the CI process (**Figure 4b-e**).

Across all the endpoints investigated in the present study, except for hERG, both mixed and expert-led CI groups outperformed individual performances, ‘all’ or ‘experts’, and the GNN models (**Figure 4a**). The most notable differences were observed for logP and permeability, with the collective SR being 20-30% higher compared to the GNN model (**Figure 4a**, shades of purple versus grey). While GNNs matched the performance of experts (unaggregated) in logP and solubility and performed worse than all CI approaches, they significantly surpassed all human responses, individual or collective, in hERG inhibition questions. For complex endpoints like solubility and logD, results from CI, individuals, and GNNs were not particularly satisfactory. Interestingly, some individuals, especially in challenging areas like hERG or logD, achieved SRs over 60%, highlighting the value of substantial expertise (**Figure 1f**, small circles).

Inspired by these results, we explored the potential of additivity between GNNs and CI, assessing their complementary strengths. We separated the answers to *incorrect* by all methods, *correct by human* (CI), *correct by GNN* or *correct by both* (**Figure 4b-e** and **S16**). For GNN the answer from the unique models was taken whereas for CI we used the log odds method applied to answers from the full cohort. Analysis showed that GNNs provided correct answers for 20% of the questions where CI failed, while CI succeeded in 32% of cases where GNN struggled. If one was able to combine GNN and CI correct answers, the overall performance would improve from 60% for the collective intelligence group to 81% with the addition of GNN, over all endpoints. The complementarity between GNNs and CI was particularly evident for solubility and hERG, where GNN would contribute 27 to 47% additional correct answers, that CI missed. Overall, a potential synergy between GNN and our collective intelligence methods would lead to an impressive SR of 87%, 81%, and 83% for hERG, solubility, and permeability, respectively. For logP, CI performs already exceptionally well while for logD the more challenging questions were missed by both CI and the GNN model (**Figure S16**).

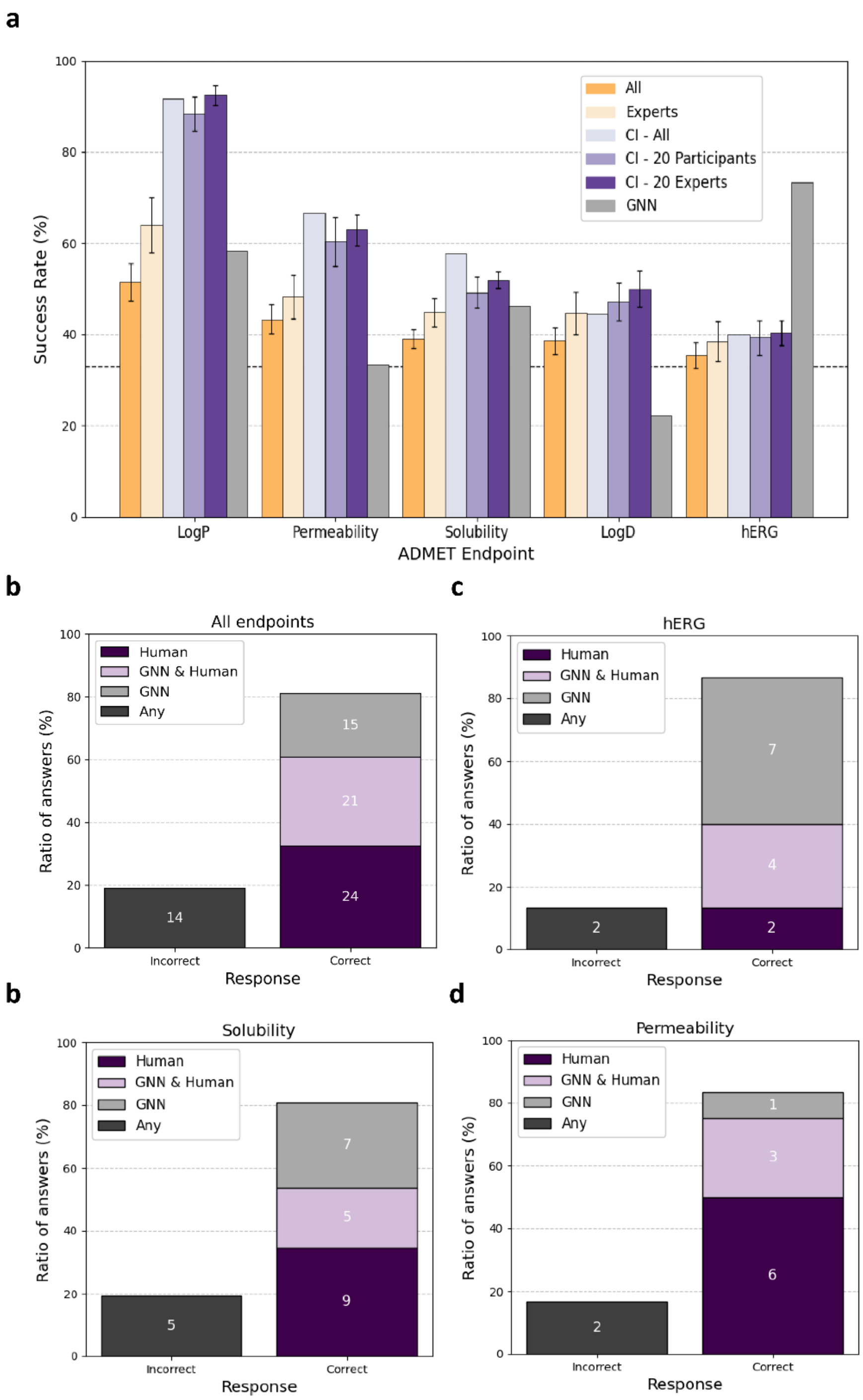


Figure 4. *Comparative benchmark of the success rate from individuals, collective, and predictive methods applied to the collective intelligence questionnaire.* **a**) Mean and standard deviation of SR for individuals, CI and a Graph Neural Network trained on public data. The expected random SR of 33% is shown as a dashed line. **b-e**) Answer success and failure ratio (y-axis) and count (numbers in boxes) for all endpoints (b), for hERG (c), for solubility (d), and for permeability (e). The answers are grouped per source, *i.e.*, human, GNN (predictive model), GNN & human, and all.

Discussion

Lead optimization campaigns are driven or are at least greatly influenced by the medicinal chemistry intuition of the project chemist leader(s). This medicinal chemistry intuition is inextricably linked to individual drug-likeness standards that depend on the chemist's experience, *'know-how'*, and bias.[35],[5],[36] Thus, characterizing a clear drug-likeness signal from medicinal chemistry intuition is a challenge.

We have presented an innovative approach to accelerate the lead optimization process that combines notions from the field of collective intelligence, medicinal chemistry, and machine learning. The responses of 92 participants to 74 medicinal chemistry multiple choice questions offered insights on the influence of expertise and confidence on the application of collective intelligence in drug discovery. It is important to emphasize that both medicinal chemistry expertise and confidence per question were by design included in our questionnaire in order to be used as parameters in the analysis and in particular in the data aggregation process.

Through ADMET optimization tasks, we observed varying success rates over self-labeled non-experts and experts in medicinal chemistry. A classification of participants based on SR revealed the superiority of teams composed of individuals with varying levels of expertise over those that lacked such variation in agreement with previous works in the cognitive science field.[8] A significant correlation between confidence levels per answer and expertise was observed, with higher confidence generally aligning with higher expertise levels. In parallel, we demonstrated that the primary determinant of SR is the level of confidence per answer expressed by participants, illustrating its importance in decision-making. Another noticeable result was the lack of correlation between SR and intermediate expertise levels. This could be due to a combination of factors: over- or under-confidence among participants, varying experience levels relative to a specific endpoint, and the inherent difficulty of the questions asked.

Performance varied across logP, logD, permeability, solubility, and hERG inhibition endpoints. Aggregation methods that account for collective intelligence significantly enhanced the success rates for tasks such as logP, permeability, and solubility, indicating the value of using such methods to address these endpoints at a project level. For example, for logP, we observe an average performance of 75%

with only 10 participants from diverse backgrounds with the log odds method that uses the information of confidence per response to aggregate the data (**Figure 2c**). In practice, this and other examples presented in the Results section (**Figure 2c** and **S12**) demonstrate the added value of collective decision-making related to at least certain ADMET endpoints.

Our results indicate that CI methods excel for endpoints involving better understood phenomena, such as logP, solubility, and permeability, but are less effective in complex areas such as hERG and logD. These challenging endpoints may require more expert input, detailed structural information, expert-focused aggregation strategies or different presentation of chemical structures, possibly including 3D information. It is likely that the format of the questionnaire itself plays a role in the participants' cognitive process, *e.g.*, a chemist would probably respond differently if the whole molecule was shown rather than only the substitution. Additionally, the brief period for responses may have limited the comprehensive evaluation of complex chemical phenomena, such as tautomeric changes and inductive or mesomeric effects, crucial for efficient optimization. A direction that could explain such cognitive phenomena and possibly further improve data aggregation is the inclusion of a timestamp per question, *i.e.*, response time. This timestamp could help elaborate the discrepancies between intermediate medicinal chemistry levels and SR, easily identify the most challenging questions, and further improve the overall SR by using for example a combination of confidence and timestamp per question as an aggregation method. Unfortunately, with the given setup we employed herein (see Methods), it was not possible to register the specific information.

The ensemble of our results demonstrates the need for tailored collective decision-making approaches in drug design, considering the varying complexities of different endpoints, the expertise of project members, and the 'expert'-'non-expert' ratio. We found that reweighting responses based on confidence improved these tasks notably. Conversely, complex endpoints like hERG and logD benefited from either an expert-dominant group or expertise-based aggregation. The study also revealed that the effectiveness of aggregation methods varied with the endpoint and group makeup. Democratic and confidence-based methods were particularly effective, especially with mixed groups of non-experts and experts. Overall, the aggregation method plays a crucial role in maximizing the performance of collective decision-making for drug design, with different methods suiting different endpoints and group compositions. This finding highlights the importance of carefully selecting aggregation methods based on the specific requirements of the task and the expertise of the participants involved. As a future direction, one might consider exploring more endpoints relevant to the lead optimization process combined with all the abovementioned aggregation methods, response time or more project-specific tasks.

The use of GNN models in our study showcased CI's ability to either match or outdo machine learning in certain domains. For logP and permeability, CI surpassed individual experts and GNNs, while in the case of hERG, the GNN model outperformed all human approaches. Our results also underscore the potential of synergy between AI and CI, particularly in complex tasks (**Figure 4b-e**). One could envisage a collective intelligence framework composed of numerous computational models, roughly equivalent in number to the participating medicinal chemists. Each model would utilize distinct descriptors or metrics, fostering a rich diversity in the decision-making process. Additionally, aggregation methods could employ iterative voting or variable weights, balancing confidence against factors like applicability domain scores. Such an approach might also benefit from transforming the typically discrete space of molecular transformations into a high-dimensional continuous decision space, thereby facilitating the identification of optimal solutions in the explored chemical series.[37] An alternative framework that would make possible such a synergy is the use of *AI medicinal chemists* that would operate within a decentralized, collaborative platform modeled in a similar manner as the Future House project (https://www.futurehouse.org/). This approach could enable a global community of researchers to contribute and refine AI-generated hypotheses, blending diverse expert insights with machine learning possibilities to enhance the discovery process.

Overall, our study demonstrates that for an effective collective intelligence-inspired drug design framework certain conditions have to be set, such as clear problem framing, appropriate aggregation methods, and a balanced team of mixed expertise, ideally comprising about 15-20 participants, to achieve significant success rates. Our results highlight CI's relevance to drug design, particularly in improving the quality of optimization proposals from a project team across various stages of drug discovery. Looking ahead, further exploration of CI for intricate tasks like hERG is essential, focusing on refining question formats and approaches to integrate structural information effectively. Another promising avenue is the hybridization of CI by blending human insights with machine learning models. This hybrid approach could leverage the strengths of both, creating a potent decision-making tool, particularly beneficial in low data regime problems. Our hope is that the CI field will continue to evolve, offering innovative and more effective solutions in the ever-complex realm of drug discovery.

Methods

Experimental Design

Population Description. This study involved a group of 92 volunteers with diverse levels of expertise in medicinal chemistry and backgrounds from analytical chemistry and crystallography to in vitro biology and data science. Before the experiment, each participant was asked to self-evaluate their expertise in medicinal chemistry on a scale from 1 (little or no experience) to 5 (expert). Throughout the present manuscript and supporting information, the results corresponding to each group are color-coded as in **Figure 1**.

Questionnaire Preparation. The experimental questions focused on late-stage lead compound optimization, targeting specific ADMET-related properties, often called endpoints in medicinal chemistry terminology, namely logP, logD, permeability, solubility, and hERG inhibition.

LogP is the logarithm of the partition coefficient (*P*) of a compound between two immiscible phases, usually octanol (as a stand-in for lipids or fats) and water (aqueous phase). It is a measure of the compound's lipophilicity and is calculated as:

$$LogP = \log\left(\frac{[C]_{octanol}}{[C]_{water}}\right),$$

Where $[C]_{oc\tan ol}$ is the concentration of the compound in octanol and $[C]_{water}$ is the concentration of the compound in water.

LogD is similar to *logP* but specifically accounts for the ionization state of a compound at a particular pH. It is the logarithm of the distribution coefficient, which quantifies the distribution of all forms (ionized and non-ionized) of the compound between the two phases, usually a Phosphate Buffer Sodium (PBS) solution (corresponding to the aqueous phase) and octanol (corresponding to the lipids phase). It is defined as:

$$LogD_{pH} = \log\left(\frac{[C]_{oc\tan ol}}{[C]_{buffer}}\right),$$

Where $[C]_{buffer}$ is the concentration of the compound in PBS buffer and $[C]_{oc\tan ol}$ is the concentration of the compound in octanol.

Permeability quantifies the rate at which a molecule crosses biological membranes, such as the intestinal epithelium. The apparent permeability (P_{app}) measured from *in vitro* assay models is calculated using the following equation:

$$P_{app} = \frac{dQ}{dt} \cdot \frac{1}{A \cdot C_0},$$

Where dQ/dt is the rate of appearance of the drug on the receiver side of the cell monolayer (in moles per time unit), A is the surface area of the cell monolayer (in square centimeters), and C_0 is the initial concentration of the drug on the donor side (in moles per volume unit).

Solubility (logS) is the maximum quantity of a solute that can dissolve in a given quantity of solvent at a specific temperature, reaching a state of thermodynamic equilibrium with the undissolved solute. The solubility of a molecule is an important factor that determines the ability to perform experimental assessment. It is often expressed in a log scale for convenience:

$$LogS = \log(C_{eq}),$$

where C_{eq} is the molar concentration of the compound in solution at equilibrium.

hERG (human Ether-à-go-go-Related Gene) refers to a gene that codes for Kv11.1 protein, the alpha subunit of a potassium ion channel in the heart, often denoted for simplicity as hERG channel. The hERG channel is crucial for the cardiac action potential's repolarization phase. Compounds that inhibit the hERG channel can prolong the QT interval on the electrocardiogram, leading to a risk of cardiac death. hERG inhibition is measured using patch-clamp electrophysiology. This method records the concentration required to inhibit 50% of the channel activity.

The format of each question consisted of a scaffold with one substitution site, accompanied by three potential modifications (**Figure S1**). The participants were instructed to select the substitution among the three options presented that in their opinion best improved a specific endpoint. By design the correct answers were, for most of the questions, significantly better than the second-best option. The questions were designed to challenge and tap into the participants' medicinal chemistry intuition without prior preparation (see also the comment below regarding the given time per question). Lead optimization tasks were gathered from the literature.[19],[20],[21],[22],[23],[24],[25],[26],[38],[39],[40],[41],[42],[43],[44],[45],[46],[47],[48],[49],[50],[51],[52],[53],[54],[55],[56],[57],[58],[59],[60],[61],[62]

Collective Intelligence Data Collection. Data collection was facilitated through PigeonHole,[63] an interactive platform that enables real-time survey. Our experiment was separated into two sessions that took place on the same day, with a break of 30mins between them. The participants used QR codes to access the questions and had 60 seconds for the first session and 30 seconds for the second session to respond. The time was adjusted during the second session after the observation that 30 seconds per question were enough for the participants. The time allowed per question was intentionally small to account for intuitive responses, however, due to technical limitations, it was not possible keep track of the response timestamp per participant. Participants were discouraged to interact and exchange with each other to avoid dilution of the results, error propagation and noise between different levels in medicinal chemistry. All participations were anonymous and labeled by the expertise level in medicinal chemistry defined by the users at the beginning of each session. The raw data collected was then standardized for subsequent analysis.

Data aggregation Methods

Different aggregation methods were tested to determine the Collective Intelligence Success Rate (CI SR), including most-frequent (also coined as ‘democratic’ in the text), confidence-weighted, expertise-weighted, confidence- and expertise-weighted, log odds, and fuzzy logic aggregation. Every method assigns a score *K* to each of the three options available (A, B, or C) and the option receiving the highest *K* score was selected as the collective answer.

Most frequent. The *most-frequent* or *'democratic'* method, also known as the mode, involves identifying the value or values that occur with the greatest frequency in a dataset. It is commonly used in scenarios where data points are categorical or discrete.

In its general form, for a dataset $X = \{x_1, x_2, \ldots, x_n\}$, which in our case is the {A, B, C}, the resulting set *C* after applying the *most-frequent* method is given by:

$$C = \{x^* \in X \parallel K_{x^*} \geq K_x, \forall x \in X\}, \tag{1}$$

where K_x represents the count of the value x for each question.

Weighting based on expertise in medicinal chemistry self-labeling. The responses are aggregated by weighing them according to the predefined expertise levels of the participants.

For each answer $x \in X = \{x_1, x_2, \ldots, x_n\}$ a score K is defined as

$$K_x = \sum_{i=1}^{l_x} w_{expertise_i}, \tag{2}$$

where $w_{expertise_i}$ is given by

$$w_{expertise_i} = \frac{expertise_i}{\sum_{j=1}^{N} expertise_j} \ , \tag{3}$$

and K_x is the score per question for each of the three options available (A, B, or C), l_x is the number of participants that answered *x* and *N* is the total number of participants. The resulting set *C* after applying the *expertise-weighted* method is given by **Equation (1)**.

Weighting based on confidence per question. The responses are aggregated by weighing them according to the confidence given in the response by the participants.

For each answer $x \ \in \ X = \{x_1, \ x_2, \ \ldots, \ x_n\}$ a score K_x is defined as

$$K_x = \sum_{i=1}^{l_x} w_{confidence_i}, \tag{4}$$

where $w_{confidence_i}$ is given by

$$w_{confidence_i} = \frac{confidence_i}{\sum_{j=1}^{N} confidence_j}, \tag{5}$$

and K_x is the score per question for each of the three options available (A, B, or C), l_x is the number of participants that answered *x* and *N* is the total number of participants. The resulting set *C* after applying the *confidence-weighted* method is again given by Equation (1).

Confidence & expertise weight. This approach combines both confidence and expertise weights for each response.

For each answer $x \ \in \ X = \{x_1, \ x_2, \ \ldots, \ x_n\}$ a score K_x is defined as

$$K_x = \sum_{i=1}^{l_x} \left(w_{expertise_i} + \ w_{confidence_i} \right), \tag{6}$$

where l_x is the number of participants that answered x, $w_{expertise_i}$ and $w_{confidence_i}$ and are given by equations (3) and (5), respectively. The resulting set C that corresponds to the 74 answers is given by Equation (1).

Log Odds. Given a set of responses where each response has an associated confidence value, the score A for each unique answer j is calculated by summing the natural logarithm of the confidence values for all instances of that answer.

Thus, the log odds score is defined as:

$$K_x = \sum_{i=1}^{l_x} ln(confidence_i), \tag{7}$$

where l_x is the number of instances for which the answer was x, and $Confidence_i$ is the confidence value for the i-th instance of $A^{log-odds}$. The answer with the highest log odds score is selected.

Fuzzy Logic Aggregation. This method employs fuzzy logic principles to aggregate data, focusing on the degree of belief (represented by confidence) in each response to determine the most likely answer.

$$K_x = \frac{\sum_{i=1}^{l_x} w_{confidence_i}}{l_x}, \tag{8}$$

where l_x is the number of instances for which the answer was x, and $confidence_i$ is the confidence value for the i-th instance of x.

Supervised & unsupervised learning applications

In this study, we employed computational methods to investigate biases in self-labelling, misconceptions regarding ADMET optimization, and the application of machine learning techniques to actively improved models using insights from collective intelligence.

Participants Map. We employed the UMAP[17] unsupervised learning algorithm implementation[64] to better compare individual participants from sessions 1 and 2 using projections in the 2D space. Training data were defined as the participants answer and confidence level. Answers were converted to numerical values (‘A’: 1, ‘B’: 2, ‘C’: 3) before scaling the data. The UMAP[65] (min_dist = 0.1, n_components = 2, n_neighbors = 15, random_state = 42) was trained without any hyperparameter optimization. For each session, the two first principal components were projected.

Chemical Space Map. The t-SNE[18] (t-distributed Stochastic Neighbor Embedding) unsupervised learning algorithm from scikit-learn[66] was used to build chemical maps from the 193 molecules comprising the CI chemical library. The ECFP4[67] fingerprints with 2048 bits were computed from all compounds before training the t-SNE (n_components = 2, perplexity = 30, random_state = 42) without any hyperparameter optimization. All compounds were then projected using the two first principal components.

Deep Learning Application

Data Gathering and Preparation. Public experimental data were sourced from three databases: OChem,[68] ChEMBL,[69] and BindingDB.[70] These datasets encompassed a range of measures such as Caco-2 apparent permeability, apparent solubility, logP, logD, and hERG pIC50. The datasets underwent a rigorous curation process to ensure quality and consistency:

- Data lacking continuous values, source information, or measured under specific conditions (*e.g.*, presence of MDR1/CYP P450 inhibitors/inducers, pH gradient conditions) were excluded.
- Data outside specified ranges for each measure (e.g., $-8 < \text{Papp} < -2$) were also removed.
- Chemical structures were then standardized through salt removal, stereochemistry elimination, aromaticity reassignment, ionization at pH 7.4, and selection of a standard tautomer.
- In case of duplicates, a single value was assigned per unique compound by keeping the median of the experimental value if the inter-laboratory (SDi) variations did not exceed 0.5 log (**Table S1**).[71]

Machine Learning Models. For the machine learning model, the datasets were divided into training (80%) and test (20%) subsets. The ChemProp GNN model[34] was trained without hyperparameters optimization and validated on the internal test set. Training parameters were defined as follows: epochs = 100, depth = 3, batch_size = 64, hidden_size = 300, and metric = rmse.

Performance Metrics To assess the performance of our models, we employed the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) (**Table S1**). R-squared measures the effectiveness of a model in explaining the variation in the dependent variable. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, with values ranging from 0 to 1. A value closer to 1 signifies a higher degree of model accuracy. RMSE evaluates the differences between predicted and actual values, emphasizing larger errors by squaring them before computing the average. This metric is particularly useful in scenarios where large deviations are especially undesirable. MAE, on the other hand, assesses the precision of a regression model. Unlike RMSE, MAE is less influenced by outliers or significant errors, as it calculates the simple average of the absolute differences between predicted and observed values.

Data Availability

The complete dataset from this study, encompassing survey responses, as well as the survey and molecules used to train and test the models, is accessible on GitHub at https://github.com/Sanofi-Public/IDD-Collective-Intelligence. We ensured ethical compliance as feedbacks from all participants remain anonymous.

Code Availability

We have made the trained models and all associated code used to from data analysis and generation publicly available under an MIT license at https://github.com/Sanofi-Public/IDD-Collective-Intelligence. For ease of integration into cheminformatics workflows, a Conda package is provided. These neural network models were developed utilizing the Chemprop library, version 1.7.0.

Acknowledgements

We express our gratitude to the researchers of Integrated Drug Discovery platform at Sanofi R&D France for their participation to our experiment and their invaluable feedback and criticism, which were essential to the success of this study. P. Llompart was funded by Sanofi and the Association Nationale de la Recherche et de la Technologie (ANRT) contract 2021-1469. The authors would also like to thank Prof. Gabriel Stoltz and Dr Rosalia Arrebola for providing feedback on the manuscript. We extend our thanks to our interns, Jianjun Tao and Margaux Dumon, for their efforts in preparing the Collective Intelligence questionnaire.

Author contributions

P. Llompart, C. Minoletti, and P. Gkeka are the main authors. Data collection, annotation process supervision, modeling and statistical analysis of results were carried out by P. Llompart, P. Mas, C. Minoletti, and P. Gkeka. Figures and tables preparation by P. Llompart under the supervision of C. Minoletti, and P. Gkeka. The first version of this article was written by P. Llompart, C. Minoletti, and P. Gkeka; K. Amaning, M. Bianciotto, B. Filoche-Rommé, Y. Foricher, D. Papin, JP. Rameau, L. Schio, M. Moussaid, G. Marcou, and A. Varnek contributed to the subsequent revisions.

Competing interests

P. Llompart, K. Amaning, M. Bianciotto, B. Filoche-Rommé, Y. Foricher, P. Mas, D. Papin, JP. Rameau, L. Schio, C. Minoletti, and P. Gkeka are Sanofi employees and may hold shares and/or stock options in the company. M. Moussaid, G. Marcou, and A. Varnek have nothing to disclose.

References

(1) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discov.* **2020**, *19* (5), 353–364. https://doi.org/10.1038/s41573-019-0050-3.

(2) Pedreira, J. G. B.; Franco, L. S.; Barreiro, E. J. Chemical Intuition in Drug Design and Discovery. *Curr. Top. Med. Chem.* **2019**, *19* (19), 1679–1693. https://doi.org/10.2174/1568026619666190620144142.

(3) Choung, O.-H.; Vianello, R.; Segler, M.; Stiefl, N.; Jiménez-Luna, J. Extracting Medicinal Chemistry Intuition via Preference Machine Learning. *Nat. Commun.* **2023**, *14* (1), 6651. https://doi.org/10.1038/s41467-023-42242-1.

(4) Gershman, S. J. How to Never Be Wrong. *Psychon. Bull. Rev.* **2019**, *26* (1), 13–28. https://doi.org/10.3758/s13423-018-1488-8.

(5) Suomala, J.; Kauttonen, J. Human’s Intuitive Mental Models as a Source of Realistic Artificial Intelligence and Engineering. *Front. Psychol.* **2022**, *13*. https://doi.org/10.3389/fpsyg.2022.873289.

(6) Gershman, S. J. *What Makes Us Smart: The Computational Logic of Human Cognition*; Princeton University Press, 2021.

(7) Woolley, A. W.; Chabris, C. F.; Pentland, A.; Hashmi, N.; Malone, T. W. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* **2010**, *330* (6004), 686–688. https://doi.org/10.1126/science.1193147.

(8) Hong, L.; Page, S. E. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. *Proc. Natl. Acad. Sci.* **2004**, *101* (46), 16385–16389. https://doi.org/10.1073/pnas.0403723101.

(9) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing Chemical Intuition in Synthesis of Metal-Organic Frameworks. *Nat. Commun.* **2019**, *10* (1), 539. https://doi.org/10.1038/s41467-019-08483-9.

(10) Duros, V.; Grizou, J.; Sharma, A.; Mehr, S. H. M.; Bubliauskas, A.; Frei, P.; Miras, H. N.; Cronin, L. Intuition-Enabled Machine Learning Beats the Competition When Joint Human-Robot Teams Perform Inorganic Chemical Experiments. *J. Chem. Inf. Model.* **2019**, *59* (6), 2664–2671. https://doi.org/10.1021/acs.jcim.9b00304.

(11) Kleffner, R.; Flatten, J.; Leaver-Fay, A.; Baker, D.; Siegel, J. B.; Khatib, F.; Cooper, S. Foldit Standalone: A Video Game-Derived Protein Structure Manipulation Interface Using Rosetta. *Bioinformatics* **2017**, *33* (17), 2765–2767. https://doi.org/10.1093/bioinformatics/btx283.

(12) Dsilva, L.; Mittal, S.; Koepnick, B.; Flatten, J.; Cooper, S.; Horowitz, S. Creating Custom Foldit Puzzles for Teaching Biochemistry. *Biochem. Mol. Biol. Educ.* **2019**, *47* (2), 133–139. https://doi.org/10.1002/bmb.21208.

(13) *Eterna*. https://eternagame.org/ (accessed 2024-04-06).

(14) Robson, J. M.; Green, A. A. Closing the Loop on Crowdsourced Science. *Proc. Natl. Acad. Sci.* **2022**, *119* (25), e2205897119. https://doi.org/10.1073/pnas.2205897119.

(15) Cincilla, G.; Masoni, S.; Blobel, J. Individual and Collective Human Intelligence in Drug Design: Evaluating the Search Strategy. *J. Cheminformatics* **2021**, *13* (1), 80. https://doi.org/10.1186/s13321-021-00556-6.

(16) Lackner, S.; Francisco, F.; Mendonça, C.; Mata, A.; Gonçalves-Sá, J. Intermediate Levels of Scientific Knowledge Are Associated with Overconfidence and Negative Attitudes towards Science. *Nat. Hum. Behav.* **2023**, *7* (9), 1490–1501. https://doi.org/10.1038/s41562-023-01677-8.

(17) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3* (29), 861. https://doi.org/10.21105/joss.00861.

(18) Cai, T. T.; Ma, R. Theoretical Foundations of T-SNE for Visualizing High-Dimensional Clustered Data. *J. Mach. Learn. Res.* **2022**, *23* (1), 301:13581-301:13634.

(19) Pevarello, P.; Brasca, M. G.; Orsini, P.; Traquandi, G.; Longo, A.; Nesi, M.; Orzi, F.; Piutti, C.; Sansonna, P.; Varasi, M.; Cameron, A.; Vulpetti, A.; Roletto, F.; Alzani, R.; Ciomei, M.; Albanese, C.; Pastori, W.; Marsiglio, A.; Pesenti, E.; Fiorentini, F.; Bischoff, J. R.; Mercurio, C. 3-Aminopyrazole Inhibitors of CDK2/Cyclin A as Antitumor Agents. 2. Lead Optimization. *J. Med. Chem.* **2005**, *48* (8), 2944–2956. https://doi.org/10.1021/jm0408870.

(20) Cui, J. J.; McTigue, M.; Nambu, M.; Tran-Dubé, M.; Pairish, M.; Shen, H.; Jia, L.; Cheng, H.; Hoffman, J.; Le, P.; Jalaie, M.; Goetz, G. H.; Ryan, K.; Grodsky, N.; Deng, Y.; Parker, M.; Timofeevski, S.; Murray, B. W.; Yamazaki, S.; Aguirre, S.; Li, Q.; Zou, H.; Christensen, J. Discovery of a Novel Class of Exquisitely Selective Mesenchymal-Epithelial Transition Factor (c-MET) Protein Kinase Inhibitors and Identification of the Clinical Candidate 2-(4-(1-(Quinolin-6-Ylmethyl)-1H-[1,2,3]Triazolo[4,5-b]Pyrazin-6-Yl)-1H-Pyrazol-1-Yl)Ethanol (PF-04217903) for the Treatment of Cancer. *J. Med. Chem.* **2012**, *55* (18), 8091–8109. https://doi.org/10.1021/jm300967g.

(21) Drews, A.; Bovens, S.; Roebrock, K.; Sunderkötter, C.; Reinhardt, D.; Schäfers, M.; Velde, A. van der; Elfringhoff, A. S.; Fabian, J.; Lehr, M. *1-(5-Carboxyindol-1-yl)propan-2-one Inhibitors of Human Cytosolic Phospholipase A2α with Reduced Lipophilicity: Synthesis, Biological Activity, Metabolic Stability, Solubility, Bioavailability, And Topical in Vivo Activity*. ACS Publications. https://doi.org/10.1021/jm1001088.

(22) Le Manach, C.; Paquet, T.; Wicht, K.; Nchinda, A. T.; Brunschwig, C.; Njoroge, M.; Gibhard,

L.; Taylor, D.; Lawrence, N.; Wittlin, S.; Eyermann, C. J.; Basarab, G. S.; Duffy, J.; Fish, P. V.; Street, L. J.; Chibale, K. Antimalarial Lead-Optimization Studies on a 2,6-Imidazopyridine Series within a Constrained Chemical Space To Circumvent Atypical Dose–Response Curves against Multidrug Resistant Parasite Strains. *J. Med. Chem.* **2018**, *61* (20), 9371–9385. https://doi.org/10.1021/acs.jmedchem.8b01333.

(23) Lin, J.; Lu, W.; Caravella, J. A.; Campbell, A. M.; Diebold, R. B.; Ericsson, A.; Fritzen, E.; Gustafson, G. R.; Lancia, D. R. Jr.; Shelekhin, T.; Wang, Z.; Castro, J.; Clarke, A.; Gotur, D.; Josephine, H. R.; Katz, M.; Diep, H.; Kershaw, M.; Yao, L.; Kauffman, G.; Hubbs, S. E.; Luke, G. P.; Toms, A. V.; Wang, L.; Bair, K. W.; Barr, K. J.; Dinsmore, C.; Walker, D.; Ashwell, S. Discovery and Optimization of Quinolinone Derivatives as Potent, Selective, and Orally Bioavailable Mutant Isocitrate Dehydrogenase 1 (mIDH1) Inhibitors. *J. Med. Chem.* **2019**, *62* (14), 6575–6596. https://doi.org/10.1021/acs.jmedchem.9b00362.

(24) Hoveyda, H. R.; Fraser, G. L.; Dutheuil, G.; El Bousmaqui, M.; Korac, J.; Lenoir, F.; Lapin, A.; Noël, S. Optimization of Novel Antagonists to the Neurokinin-3 Receptor for the Treatment of Sex-Hormone Disorders (Part II). *ACS Med. Chem. Lett.* **2015**, *6* (7), 736–740. https://doi.org/10.1021/acsmedchemlett.5b00117.

(25) Richter, H. G. F.; Benson, G. M.; Bleicher, K. H.; Blum, D.; Chaput, E.; Clemann, N.; Feng, S.; Gardes, C.; Grether, U.; Hartman, P.; Kuhn, B.; Martin, R. E.; Plancher, J.-M.; Rudolph, M. G.; Schuler, F.; Taylor, S. Optimization of a Novel Class of Benzimidazole-Based Farnesoid X Receptor (FXR) Agonists to Improve Physicochemical and ADME Properties. *Bioorg. Med. Chem. Lett.* **2011**, *21* (4), 1134–1140. https://doi.org/10.1016/j.bmcl.2010.12.123.

(26) Koda, Y.; Sato, S.; Yamamoto, H.; Niwa, H.; Watanabe, H.; Watanabe, C.; Sato, T.; Nakamura, K.; Tanaka, A.; Shirouzu, M.; Honma, T.; Fukami, T.; Koyama, H.; Umehara, T. Design and Synthesis of Tranylcypromine-Derived LSD1 Inhibitors with Improved hERG and Microsomal Stability Profiles. *ACS Med. Chem. Lett.* **2022**, *13* (5), 848–854. https://doi.org/10.1021/acsmedchemlett.2c00120.

(27) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure. *Adv. Drug Deliv. Rev.* **2002**, *54* (3), 355–366. https://doi.org/10.1016/S0169-409X(02)00008-X.

(28) Romanelli, M. N.; Manetti, D.; Braconi, L.; Dei, S.; Gabellini, A.; Teodori, E. The Piperazine Scaffold for Novel Drug Discovery Efforts: The Evidence to Date. *Expert Opin. Drug Discov.* **2022**.

(29) Kumari, A.; Singh, R. K. Morpholine as Ubiquitous Pharmacophore in Medicinal Chemistry: Deep Insight into the Structure-Activity Relationship (SAR). *Bioorganic Chem.* **2020**, *96*, 103578. https://doi.org/10.1016/j.bioorg.2020.103578.

(30) Talele, T. T. The "Cyclopropyl Fragment" Is a Versatile Player That Frequently Appears in Preclinical/Clinical Drug Molecules. *J. Med. Chem.* **2016**, *59* (19), 8712–8756. https://doi.org/10.1021/acs.jmedchem.6b00472.

(31) Comer, J. E. A. High-Throughput Measurement of Log D and pKa. In *Drug Bioavailability*; John Wiley & Sons, Ltd, 2003; pp 21–45. https://doi.org/10.1002/3527601473.ch2.

(32) Landry, M. L.; Crawford, J. J. LogD Contributions of Substituents Commonly Used in Medicinal Chemistry. *ACS Med. Chem. Lett.* **2020**, *11* (1), 72–76. https://doi.org/10.1021/acsmedchemlett.9b00489.

(33) Kalyaanamoorthy, S.; Barakat, K. H. Binding Modes of hERG Blockers: An Unsolved Mystery in the Drug Design Arena. *Expert Opin. Drug Discov.* **2018**, *13* (3), 207–210. https://doi.org/10.1080/17460441.2018.1418319.

(34) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.

(35) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47* (20), 4891–4896. https://doi.org/10.1021/jm049740z.

(36) Jolly, E.; Chang, L. J. The Flatland Fallacy: Moving Beyond Low–Dimensional Thinking. *Top. Cogn. Sci.* **2019**, *11* (2), 433–454. https://doi.org/10.1111/tops.12404.

(37) Garg, N.; Kamble, V.; Goel, A.; Marn, D.; Munagala, K. Iterative Local Voting for Collective Decision-Making in Continuous Spaces. *J. Artif. Intell. Res.* **2019**, *64*, 315–355. https://doi.org/10.1613/jair.1.11358.

(38) Bradbury, R. H.; Callis, R.; Carr, G. R.; Chen, H.; Clark, E.; Feron, L.; Glossop, S.; Graham, M. A.; Hattersley, M.; Jones, C.; Lamont, S. G.; Ouvry, G.; Patel, A.; Patel, J.; Rabow, A. A.; Roberts, C. A.; Stokes, S.; Stratton, N.; Walker, G. E.; Ward, L.; Whalley, D.; Whittaker, D.; Wrigley, G.; Waring, M. J. Optimization of a Series of Bivalent Triazolopyridazine Based Bromodomain and Extraterminal Inhibitors: The Discovery of (3R)-4-[2-[4-[1-(3-Methoxy-[1,2,4]Triazolo[4,3-b]Pyridazin-6-Yl)-4-Piperidyl]Phenoxy]Ethyl]-1,3-Dimethyl-Piperazin-2-One (AZD5153). *J. Med. Chem.* **2016**, *59* (17), 7801–7817. https://doi.org/10.1021/acs.jmedchem.6b00070.

(39) Bovens, S.; Schulze Elfringhoff, A.; Kaptur, M.; Reinhardt, D.; Schäfers, M.; Lehr, M. 1-(5-Carboxyindol-1-Yl)Propan-2-One Inhibitors of Human Cytosolic Phospholipase A2α: Effect of Substituents in Position 3 of the Indole Scaffold on Inhibitory Potency, Metabolic Stability, Solubility, and Bioavailability. *J. Med. Chem.* **2010**, *53* (23), 8298–8308. https://doi.org/10.1021/jm101094p.

(40) Ishikawa, M.; Hashimoto, Y. Improvement in Aqueous Solubility in Small Molecule Drug Discovery Programs by Disruption of Molecular Planarity and Symmetry. *J. Med. Chem.* **2011**, *54* (6), 1539–1554. https://doi.org/10.1021/jm101356p.

(41) Couturier, C.; Lair, C.; Pellet, A.; Upton, A.; Kaneko, T.; Perron, C.; Cogo, E.; Menegotto, J.; Bauer, A.; Scheiper, B.; Lagrange, S.; Bacqué, E. Identification and Optimization of a New Series of Anti-Tubercular Quinazolinones. *Bioorg. Med. Chem. Lett.* **2016**, *26* (21), 5290–5299. https://doi.org/10.1016/j.bmcl.2016.09.043.

(42) Hanrahan, P.; Bell, J.; Bottomley, G.; Bradley, S.; Clarke, P.; Curtis, E.; Davis, S.; Dawson, G.; Horswill, J.; Keily, J.; Moore, G.; Rasamison, C.; Bloxham, J. Substituted Azaquinazolinones as

Modulators of GHSr-1a for the Treatment of Type II Diabetes and Obesity. *Bioorg. Med. Chem. Lett.* **2012**, *22* (6), 2271–2278. https://doi.org/10.1016/j.bmcl.2012.01.078.

(43) Kuttruff, C. A.; Ferrara, M.; Bretschneider, T.; Hoerer, S.; Handschuh, S.; Nosse, B.; Romig, H.; Nicklin, P.; Roth, G. J. Discovery of BI-2545: A Novel Autotaxin Inhibitor That Significantly Reduces LPA Levels in Vivo. *ACS Med. Chem. Lett.* **2017**, *8* (12), 1252–1257. https://doi.org/10.1021/acsmedchemlett.7b00312.

(44) Panchaud, P.; Bruyère, T.; Blumstein, A.-C.; Bur, D.; Chambovey, A.; Ertel, E. A.; Gude, M.; Hubschwerlen, C.; Jacob, L.; Kimmerlin, T.; Pfeifer, T.; Prade, L.; Seiler, P.; Ritz, D.; Rueedi, G. Discovery and Optimization of Isoquinoline Ethyl Ureas as Antibacterial Agents. *J. Med. Chem.* **2017**, *60* (9), 3755–3775. https://doi.org/10.1021/acs.jmedchem.6b01834.

(45) Hameed P, S.; Patil, V.; Solapure, S.; Sharma, U.; Madhavapeddi, P.; Raichurkar, A.; Chinnapattu, M.; Manjrekar, P.; Shanbhag, G.; Puttur, J.; Shinde, V.; Menasinakai, S.; Rudrapatana, S.; Achar, V.; Awasthy, D.; Nandishaiah, R.; Humnabadkar, V.; Ghosh, A.; Narayan, C.; Ramya, V. K.; Kaur, P.; Sharma, S.; Werngren, J.; Hoffner, S.; Panduga, V.; Kumar, C. N. N.; Reddy, J.; Kumar KN, M.; Ganguly, S.; Bharath, S.; Bheemarao, U.; Mukherjee, K.; Arora, U.; Gaonkar, S.; Coulson, M.; Waterson, D.; Sambandamurthy, V. K.; de Sousa, S. M. Novel N-Linked Aminopiperidine-Based Gyrase Inhibitors with Improved hERG and in Vivo Efficacy against Mycobacterium Tuberculosis. *J. Med. Chem.* **2014**, *57* (11), 4889–4905. https://doi.org/10.1021/jm500432n.

(46) Subbaiah, M. A. M.; Meanwell, N. A. Bioisosteres of the Phenyl Ring: Recent Strategic Applications in Lead Optimization and Drug Design. *J. Med. Chem.* **2021**, *64* (19), 14046–14128. https://doi.org/10.1021/acs.jmedchem.1c01215.

(47) Huang, S.-C.; Adhikari, S.; Afroze, R.; Brewer, K.; Calderwood, E. F.; Chouitar, J.; England, D. B.; Fisher, C.; Galvin, K. M.; Gaulin, J.; Greenspan, P. D.; Harrison, S. J.; Kim, M.-S.; Langston, S. P.; Ma, L.-T.; Menon, S.; Mizutani, H.; Rezaei, M.; Smith, M. D.; Zhang, D. M.; Gould, A. E. Optimization of Tetrahydronaphthalene Inhibitors of Raf with Selectivity over hERG. *Bioorg. Med. Chem. Lett.* **2016**, *26* (4), 1156–1160. https://doi.org/10.1016/j.bmcl.2016.01.049.

(48) Kazmierski, W. M.; Anderson, D. L.; Aquino, C.; Chauder, B. A.; Duan, M.; Ferris, R.; Kenakin, T.; Koble, C. S.; Lang, D. G.; Mcintyre, M. S.; Peckham, J.; Watson, C.; Wheelan, P.; Spaltenstein, A.; Wire, M. B.; Svolto, A.; Youngman, M. Novel 4,4-Disubstituted Piperidine-Based C–C Chemokine Receptor-5 Inhibitors with High Potency against Human Immunodeficiency Virus-1 and an Improved Human Ether-a-Go-Go Related Gene (hERG) Profile. *J. Med. Chem.* **2011**, *54* (11), 3756–3767. https://doi.org/10.1021/jm200279v.

(49) Dorado, T. E.; de León, P.; Begum, A.; Liu, H.; Chen, D.; Rajeshkumar, N. V.; Rey-Rodriguez, R.; Hoareau-Aveilla, C.; Alcouffe, C.; Laiho, M.; Barrow, J. C. Discovery and Evaluation of Novel Angular Fused Pyridoquinazolinonecarboxamides as RNA Polymerase I Inhibitors. *ACS Med. Chem. Lett.* **2022**, *13* (4), 608–614. https://doi.org/10.1021/acsmedchemlett.1c00660.

(50) Rynearson, K. D.; Buckle, R. N.; Barnes, K. D.; Herr, R. J.; Mayhew, N. J.; Paquette, W. D.; Sakwa, S. A.; Nguyen, P. D.; Johnson, G.; Tanzi, R. E.; Wagner, S. L. Design and Synthesis of Aminothiazole Modulators of the Gamma-Secretase Enzyme. *Bioorg. Med. Chem. Lett.* **2016**, *26* (16), 3928–3937. https://doi.org/10.1016/j.bmcl.2016.07.011.

(51) Vijay Kumar, D.; Hoarau, C.; Bursavich, M.; Slattum, P.; Gerrish, D.; Yager, K.; Saunders, M.; Shenderovich, M.; Roth, B. L.; McKinnon, R.; Chan, A.; Cimbora, D. M.; Bradford, C.; Reeves, L.; Patton, S.; Papac, D. I.; Williams, B. L.; Carlson, R. O. Lead Optimization of Purine Based Orally Bioavailable Mps1 (TTK) Inhibitors. *Bioorg. Med. Chem. Lett.* **2012**, *22* (13), 4377–4385. https://doi.org/10.1016/j.bmcl.2012.04.131.

(52) Harnden, A. C.; Davis, O. A.; Box, G. M.; Hayes, A.; Johnson, L. D.; Henley, A. T.; de Haven Brandon, A. K.; Valenti, M.; Cheung, K.-M. J.; Brennan, A.; Huckvale, R.; Pierrat, O. A.; Talbot, R.; Bright, M. D.; Akpinar, H. A.; Miller, D. S. J.; Tarantino, D.; Gowan, S.; de Klerk, S.; McAndrew, P. C.; Le Bihan, Y.-V.; Meniconi, M.; Burke, R.; Kirkin, V.; van Montfort, R. L. M.; Raynaud, F. I.; Rossanese, O. W.; Bellenie, B. R.; Hoelder, S. Discovery of an In Vivo Chemical Probe for BCL6 Inhibition by Optimization of Tricyclic Quinolinones. *J. Med. Chem.* **2023**, *66* (8), 5892–5906. https://doi.org/10.1021/acs.jmedchem.3c00155.

(53) Nair, A. G.; Wong, M. K. C.; Shu, Y.; Jiang, Y.; Jenh, C.-H.; Kim, S. H.; Yang, D.-Y.; Zeng, Q.; Shao, Y.; Zawacki, L. G.; Duo, J.; McGuinness, B. F.; Carroll, C. D.; Hobbs, D. W.; Shih, N.-Y.; Rosenblum, S. B.; Kozlowski, J. A. IV. Discovery of CXCR3 Antagonists Substituted with Heterocycles as Amide Surrogates: Improved PK, hERG and Metabolic Profiles. *Bioorg. Med. Chem. Lett.* **2014**, *24* (4), 1085–1088. https://doi.org/10.1016/j.bmcl.2014.01.009.

(54) Wilson, D. M.; Apps, J.; Bailey, N.; Bamford, M. J.; Beresford, I. J.; Brackenborough, K.; Briggs, M. A.; Brough, S.; Calver, A. R.; Crook, B.; Davis, R. K.; Davis, R. P.; Davis, S.; Dean, D. K.; Harris, L.; Heslop, T.; Holland, V.; Jeffrey, P.; Panchal, T. A.; Parr, C. A.; Quashie, N.; Schogger, J.; Sehmi, S. S.; Stean, T. O.; Steadman, J. G. A.; Trail, B.; Wald, J.; Worby, A.; Takle, A. K.; Witherington, J.; Medhurst, A. D. Identification of Clinical Candidates from the Benzazepine Class of Histamine H3 Receptor Antagonists. *Bioorg. Med. Chem. Lett.* **2013**, *23* (24), 6890–6896. https://doi.org/10.1016/j.bmcl.2013.09.090.

(55) Rolt, A.; Talley, D. C.; Park, S. B.; Hu, Z.; Dulcey, A.; Ma, C.; Irvin, P.; Leek, M.; Wang, A. Q.; Stachulski, A. V.; Xu, X.; Southall, N.; Ferrer, M.; Liang, T. J.; Marugan, J. J. Discovery and Optimization of a 4-Aminopiperidine Scaffold for Inhibition of Hepatitis C Virus Assembly. *J. Med. Chem.* **2021**, *64* (13), 9431–9443. https://doi.org/10.1021/acs.jmedchem.1c00696.

(56) Kobayashi, D.; Kuraoka, E.; Hayashi, J.; Yasuda, T.; Kohmura, Y.; Denda, M.; Harada, N.; Inagaki, N.; Otaka, A. S-Protected Cysteine Sulfoxide-Enabled Tryptophan-Selective Modification with Application to Peptide Lipidation. *ACS Med. Chem. Lett.* **2022**, *13* (7), 1125–1130. https://doi.org/10.1021/acsmedchemlett.2c00161.

(57) Woodring, J. L.; Bachovchin, K. A.; Brady, K. G.; Gallerstein, M. F.; Erath, J.; Tanghe, S.; Leed, S. E.; Rodriguez, A.; Mensa-Wilmot, K.; Sciotti, R. J.; Pollastri, M. P. Optimization of Physicochemical Properties for 4-Anilinoquinazoline Inhibitors of Trypanosome Proliferation. *Eur. J. Med. Chem.* **2017**, *141*, 446–459. https://doi.org/10.1016/j.ejmech.2017.10.007.

(58) Lee, W.; Crawford, J. J.; Aliagas, I.; Murray, L. J.; Tay, S.; Wang, W.; Heise, C. E.; Hoeflich, K. P.; La, H.; Mathieu, S.; Mintzer, R.; Ramaswamy, S.; Rouge, L.; Rudolph, J. Synthesis and

Evaluation of a Series of 4-Azaindole-Containing P21-Activated Kinase-1 Inhibitors. *Bioorg. Med. Chem. Lett.* **2016**, *26* (15), 3518–3524. https://doi.org/10.1016/j.bmcl.2016.06.031.

(59) Kuriwaki, I.; Kameda, M.; Iikubo, K.; Hisamichi, H.; Kawamoto, Y.; Kikuchi, S.; Moritomo, H.; Terasaka, T.; Iwai, Y.; Noda, A.; Tomiyama, H.; Kikuchi, A.; Hirano, M. Discovery of ASP5878: Synthesis and Structure–Activity Relationships of Pyrimidine Derivatives as Pan-FGFRs Inhibitors with Improved Metabolic Stability and Suppressed *h*ERG Channel Inhibitory Activity. *Bioorg. Med. Chem.* **2022**, *59*, 116657. https://doi.org/10.1016/j.bmc.2022.116657.

(60) Goldberg, F. W.; Ting, A. K. T.; Beattie, D.; Lamont, G. M.; Fallan, C.; Finlay, M. R. V.; Williamson, B.; Schimpl, M.; Harmer, A. R.; Adeyemi, O. B.; Nordell, P.; Cronin, A. S.; Vazquez-Chantada, M.; Barratt, D.; Ramos-Montoya, A.; Cadogan, E. B.; Davies, B. R. Optimization of hERG and Pharmacokinetic Properties for Basic Dihydro-8H-Purin-8-One Inhibitors of DNA-PK. *ACS Med. Chem. Lett.* **2022**, *13* (8), 1295–1301. https://doi.org/10.1021/acsmedchemlett.2c00172.

(61) Reichard, H. A.; Schiffer, H. H.; Monenschein, H.; Atienza, J. M.; Corbett, G.; Skaggs, A. W.; Collia, D. R.; Ray, W. J.; Serrats, J.; Bliesath, J.; Kaushal, N.; Lam, B. P.; Amador-Arjona, A.; Rahbaek, L.; McConn, D. J.; Mulligan, V. J.; Brice, N.; Gaskin, P. L. R.; Cilia, J.; Hitchcock, S. Discovery of TAK-041: A Potent and Selective GPR139 Agonist Explored for the Treatment of Negative Symptoms Associated with Schizophrenia. *J. Med. Chem.* **2021**, *64* (15), 11527–11542. https://doi.org/10.1021/acs.jmedchem.1c00820.

(62) Large, J. M.; Osborne, S. A.; Smiljanic-Hurley, E.; Ansell, K. H.; Jones, H. M.; Taylor, D. L.; Clough, B.; Green, J. L.; Holder, A. A. Imidazopyridazines as Potent Inhibitors of *Plasmodium Falciparum* Calcium-Dependent Protein Kinase 1 (*Pf*CDPK1): Preparation and Evaluation of Pyrazole Linked Analogues. *Bioorg. Med. Chem. Lett.* **2013**, *23* (21), 6019–6024. https://doi.org/10.1016/j.bmcl.2013.08.010.

(63) PigeonLab. *Engage your audience with Pigeonhole Live*. https://pigeonholelive.com/ (accessed 2024-04-07).

(64) McInnes, L. Lmcinnes/Umap, 2024. https://github.com/lmcinnes/umap (accessed 2024-04-07).

(65) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv September 17, 2020. https://doi.org/10.48550/arXiv.1802.03426.

(66) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *Mach. Learn. PYTHON*.

(67) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(68) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided Mol. Des.* **2011**, *25* (6), 533–554. https://doi.org/10.1007/s10822-011-9440-2.

(69) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

(70) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A Web-Accessible Molecular Recognition Database. *Comb. Chem. High Throughput Screen.* **2001**, *4* (8), 719–725. https://doi.org/10.2174/1386207013330670.

(71) Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M. Á. Reliable Prediction of Caco-2

Permeability by Supervised Recursive Machine Learning Approaches. *Pharmaceutics* **2022**, *14* (10), 1998. https://doi.org/10.3390/pharmaceutics14101998.

Supplementary Information:

Harnessing Medicinal Chemical Intuition from Collective Intelligence

Pierre Llompart[1, 2, *], Kwame Amaning[1], Marc Bianciotto[1], Bruno Filoche-Rommé[1], Yann Foricher[1], Pablo Mas[1,3], David Papin[1], Jean-Philippe Rameau[1], Laurent Schio[1], Gilles Marcou[2], Alexandre Varnek[2], Mehdi Moussaid[4,5], Claire Minoletti[1, *], Paraskevi Gkeka[1, *]

[1]Integrated Drug Discovery, Sanofi, Vitry-sur-Seine, France

[2]Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg, France

[3] Theoretical Chemistry Department, École Normale Supérieure, Paris, France

[4]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[5]School of Collective Intelligence, Université Mohammed VI Polytechnique, Rabat, Morocco

*Corresponding authors: Paraskevi.Gkeka@sanofi.com, Pierre.Llompart@sanofi.com, Claire.Minoletti@sanofi.com

Keywords: medicinal chemistry intuition, collective intelligence, ADMET, drug design, GNN, lead optimization

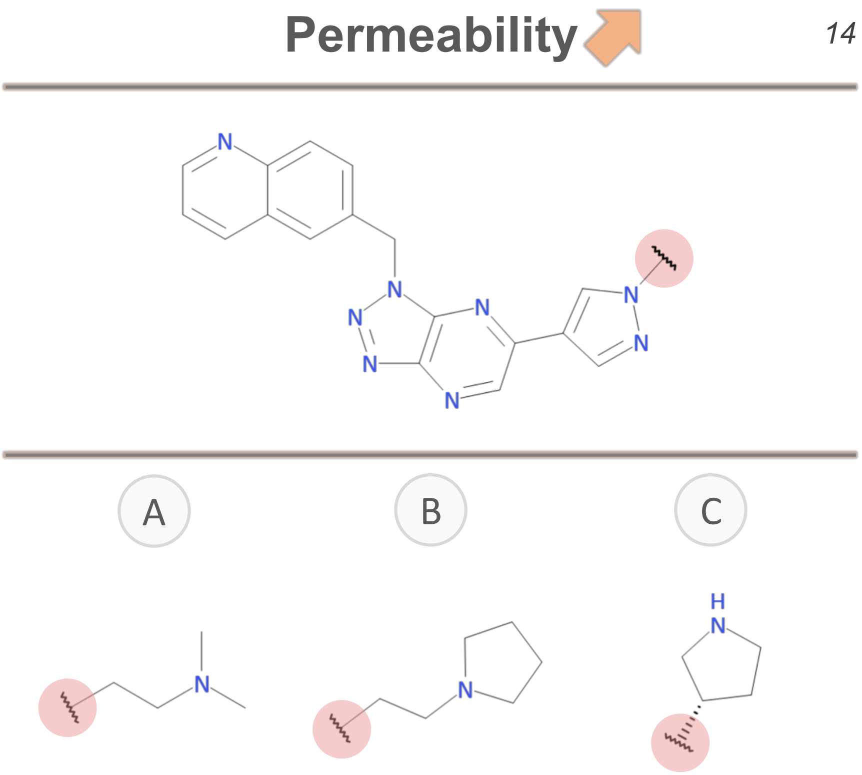


Figure S1: *Example of the way the questions were presented to the participants.* Each question had a title that corresponded to the endpoint of interest (top), one scaffold with an R-group substitution point to be replaced (middle) and three possible substituents (bottom).

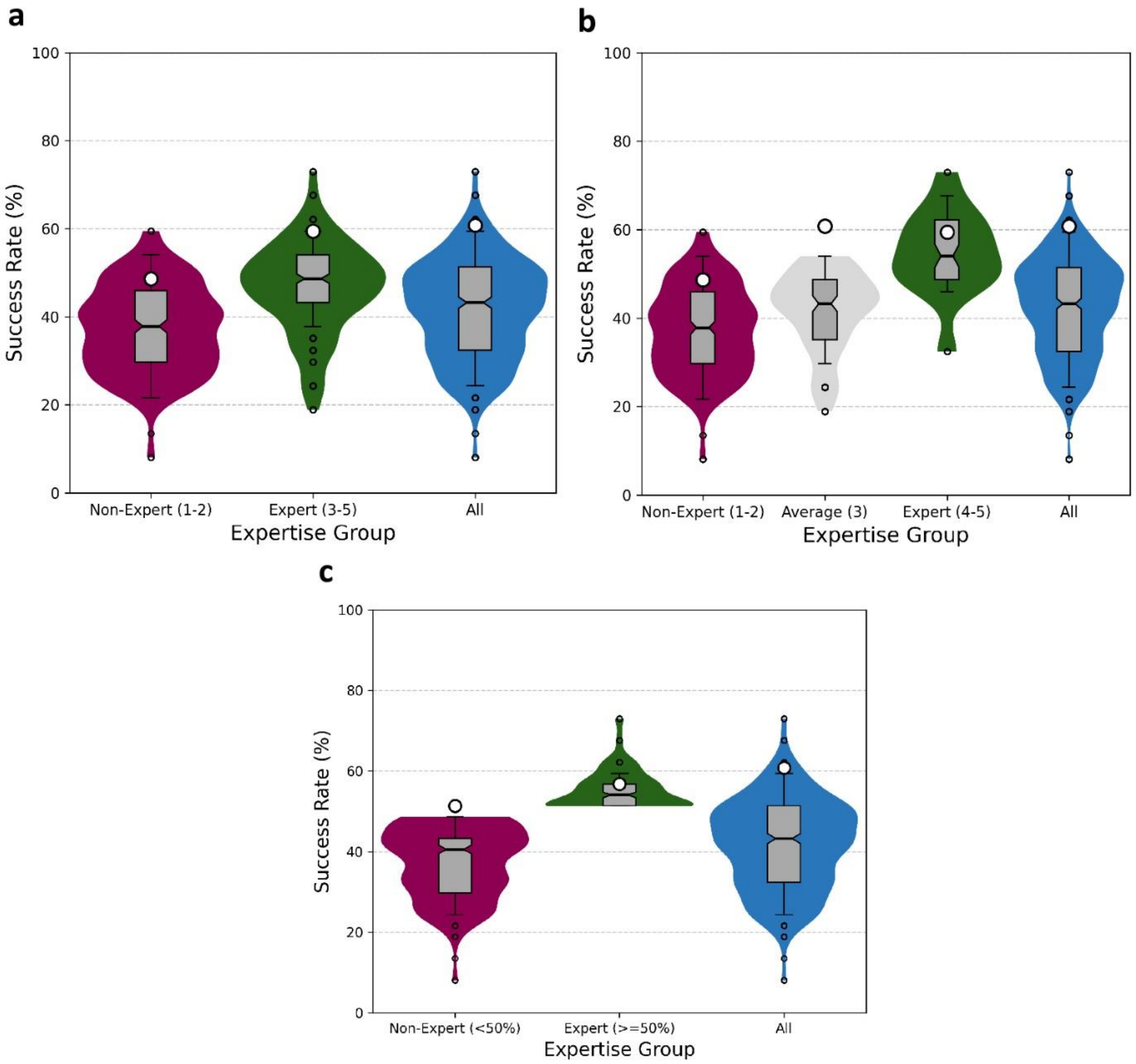


Figure S2: *Violin plots representing the success rate by expertise level in medicinal chemistry*. **a)** SR for groups 1-2 (low or no background), 3-5 (experts) and all the participants. The median is shown as a horizontal line across the thinnest part of the grey boxes. Alongside the boxes, error bars extend from the median line to cover the interquartile range. The collective or democratic SR are shown as white-filled circles. The outliers per group are depicted as small circles. **b)** SR for groups 1-2 (low or no background), 3 (averaged and mixed level), 4-5 (experts) and all the participants. **c)** SR by non-experts, *i.e.*, participants with personal SR less than 50%, and experts, *i.e.*, individuals with SR more than 50%. The violin plot of the SR of all participants is shown again here as a guide for the eye.

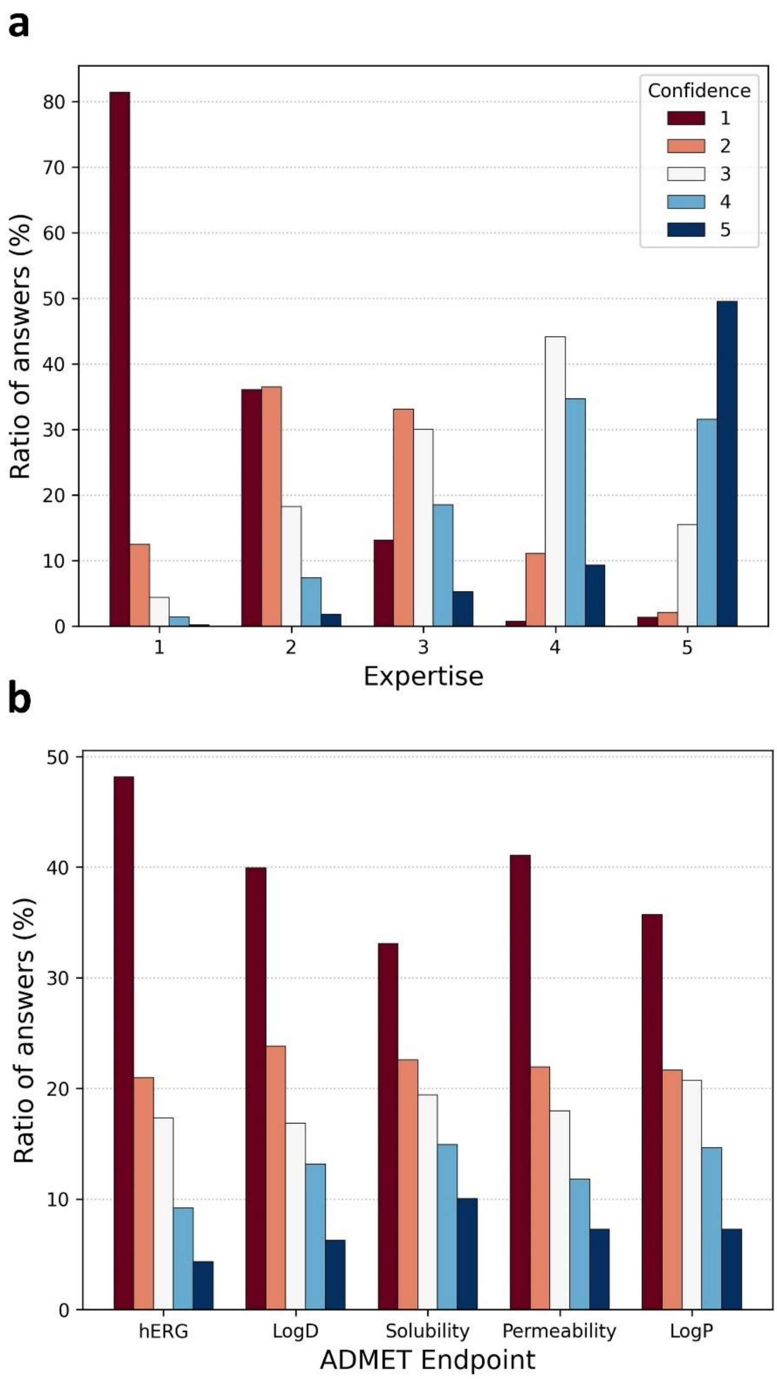


Figure S3: *Distribution of the proportion of answer per confidence level in function of the ADMET endpoint and self-labeled expertise level.* a) Bar plot of the ratio of answers per expertise level grouped by confidence from low (red) to high (blue). b) Bar plot of the ratio of answers per ADMET endpoint.

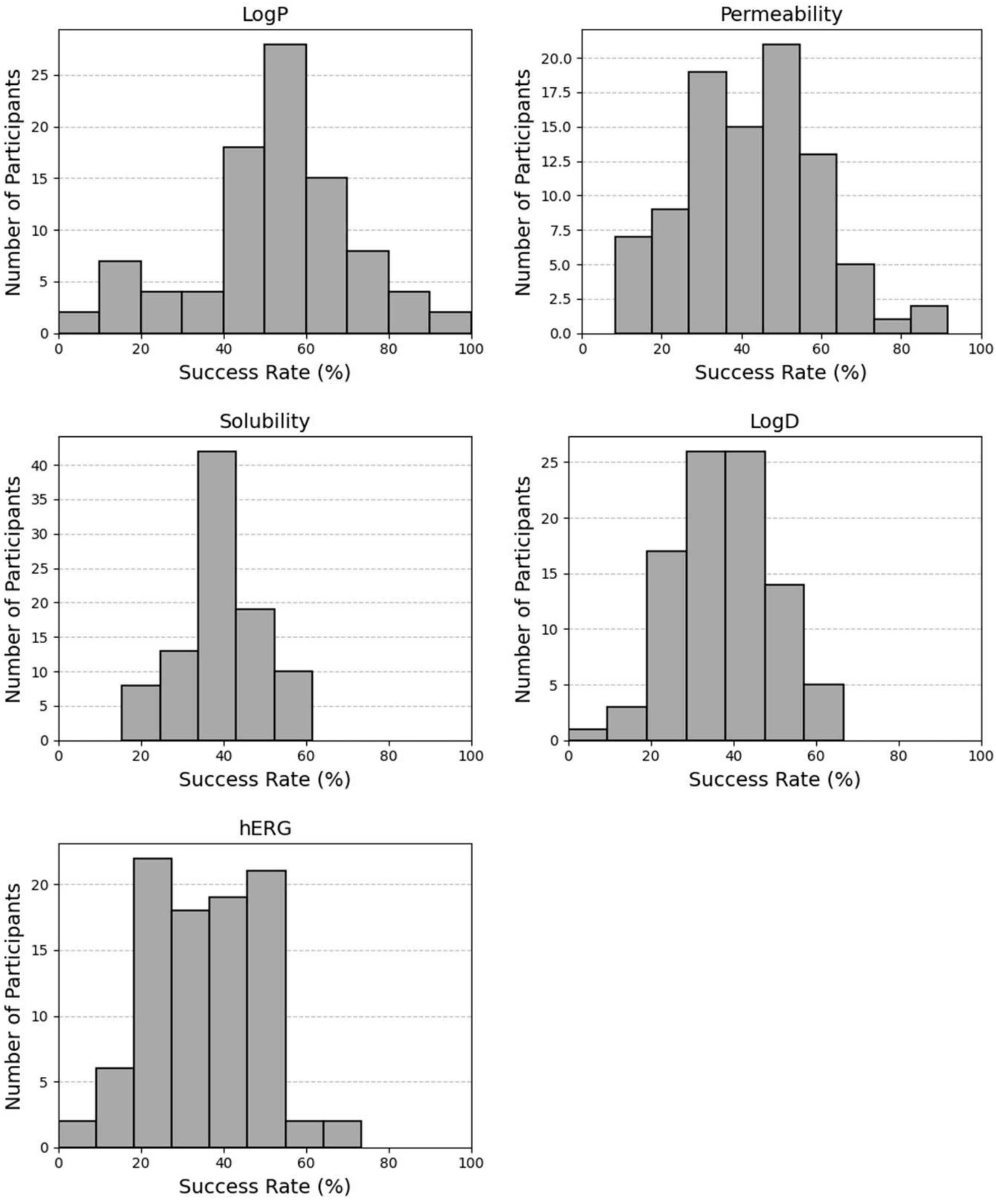


Figure S4: *Success rate distribution for all participants per ADMET endpoint.*

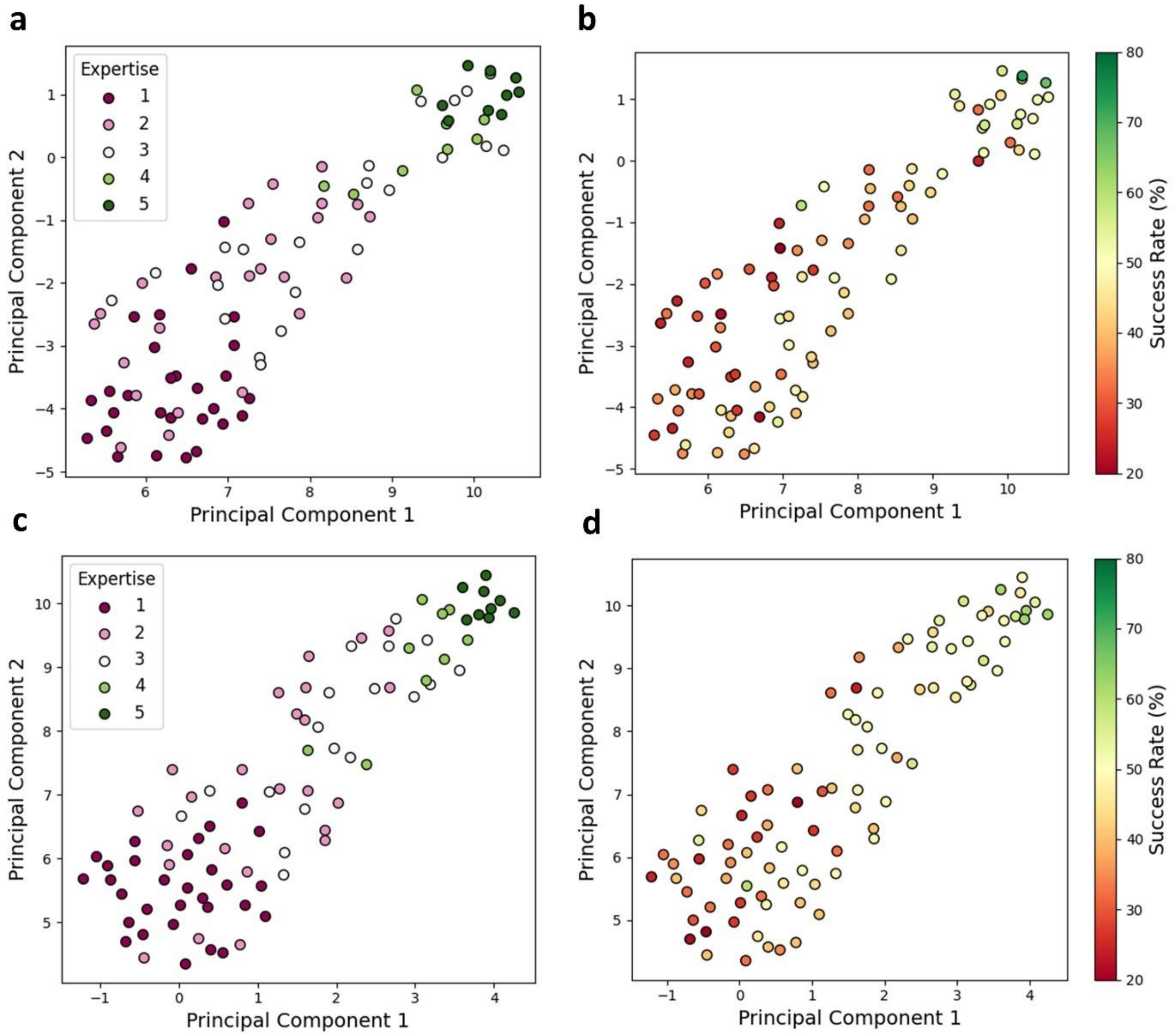


Figure S5: *UMAP of the participant space per session explored using the expertise level and the success rate.* Session one colored by a) expertise level and b) success rate. Session two colored by c) expertise level and d) success rate.

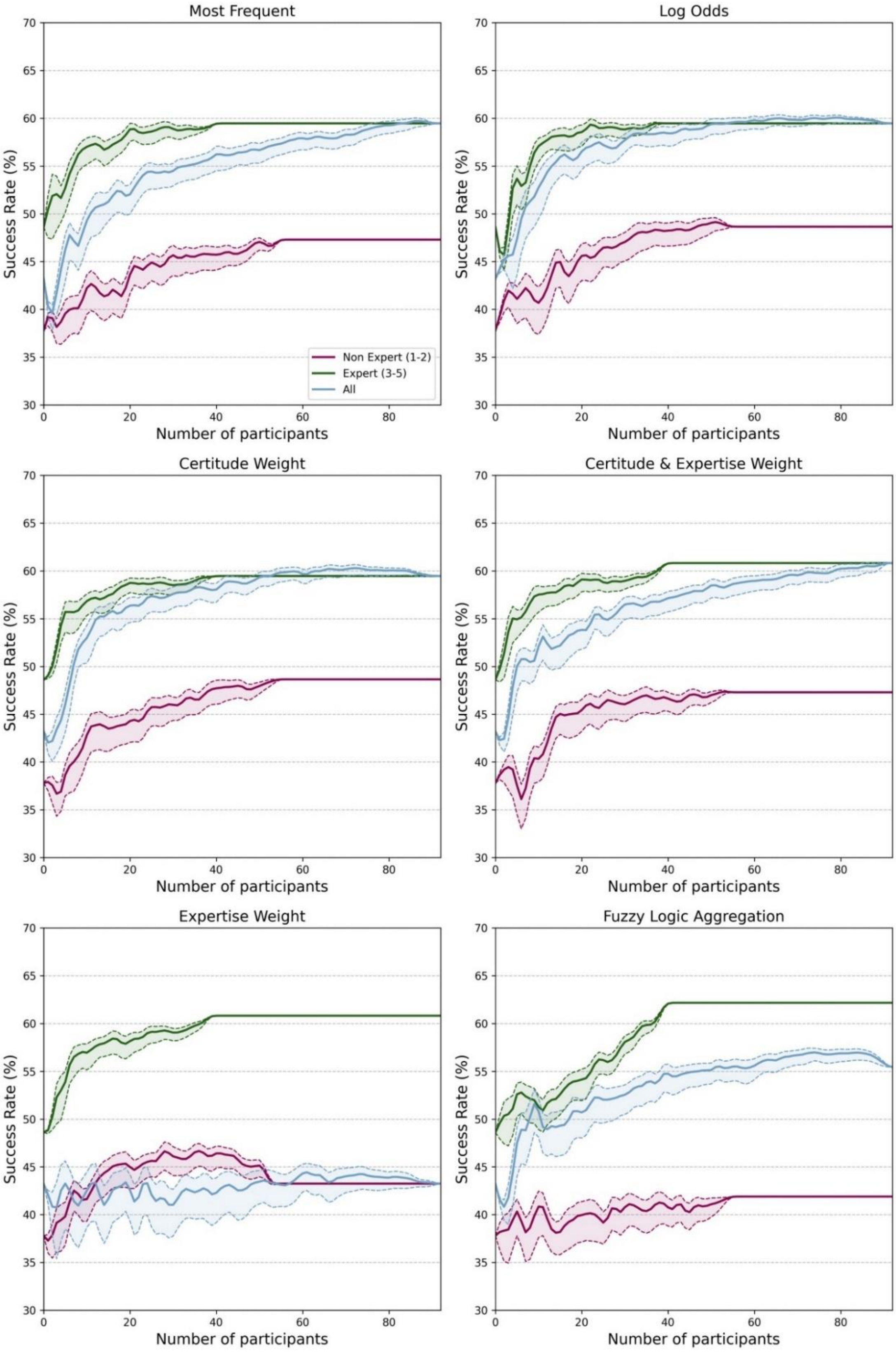


Figure S6: *Evolution of the collective success rate as a function of the number of participants in the population per aggregation method.* The collective SR is denoted per expertise group, from non-expert (1-2), expert (3-5), and all participants.

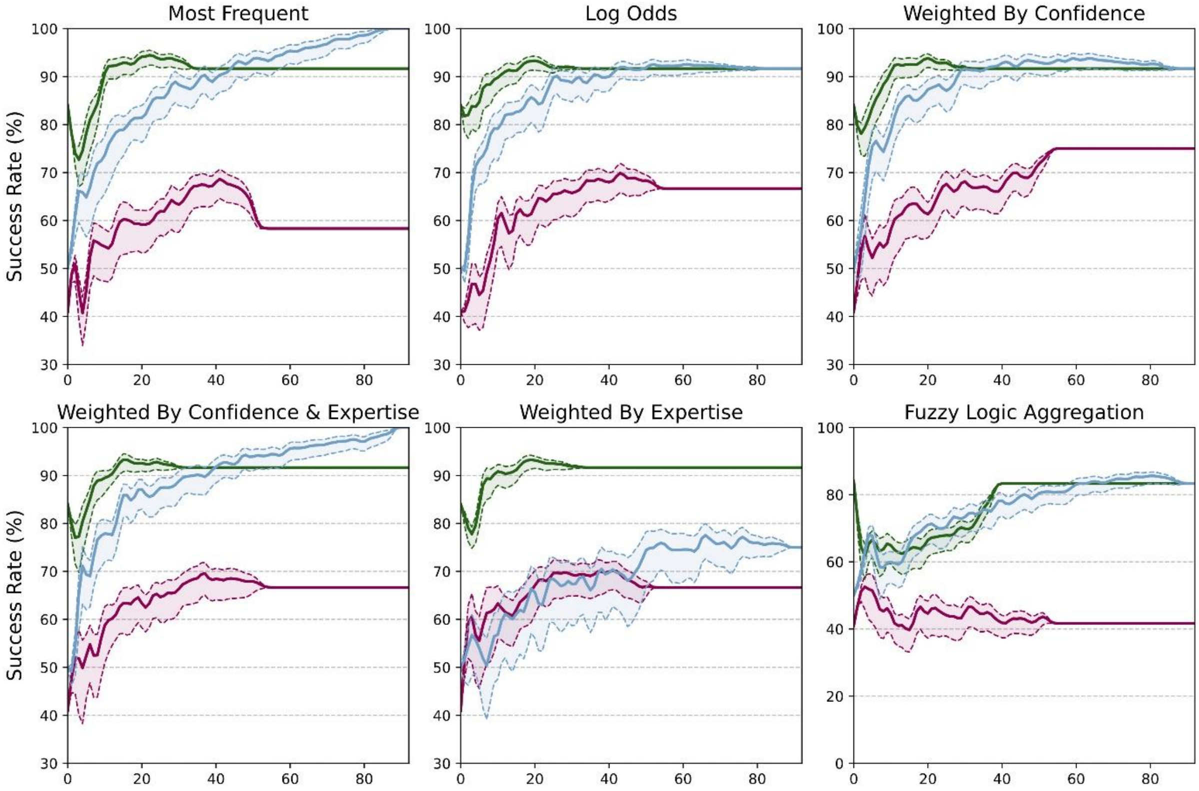


Figure S7: *Evolution of the collective success rate as a function of the number of participants in the population per aggregation method for the LogP endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2), expert (3-5), and all participants. Groups are colored as in Figure S6.

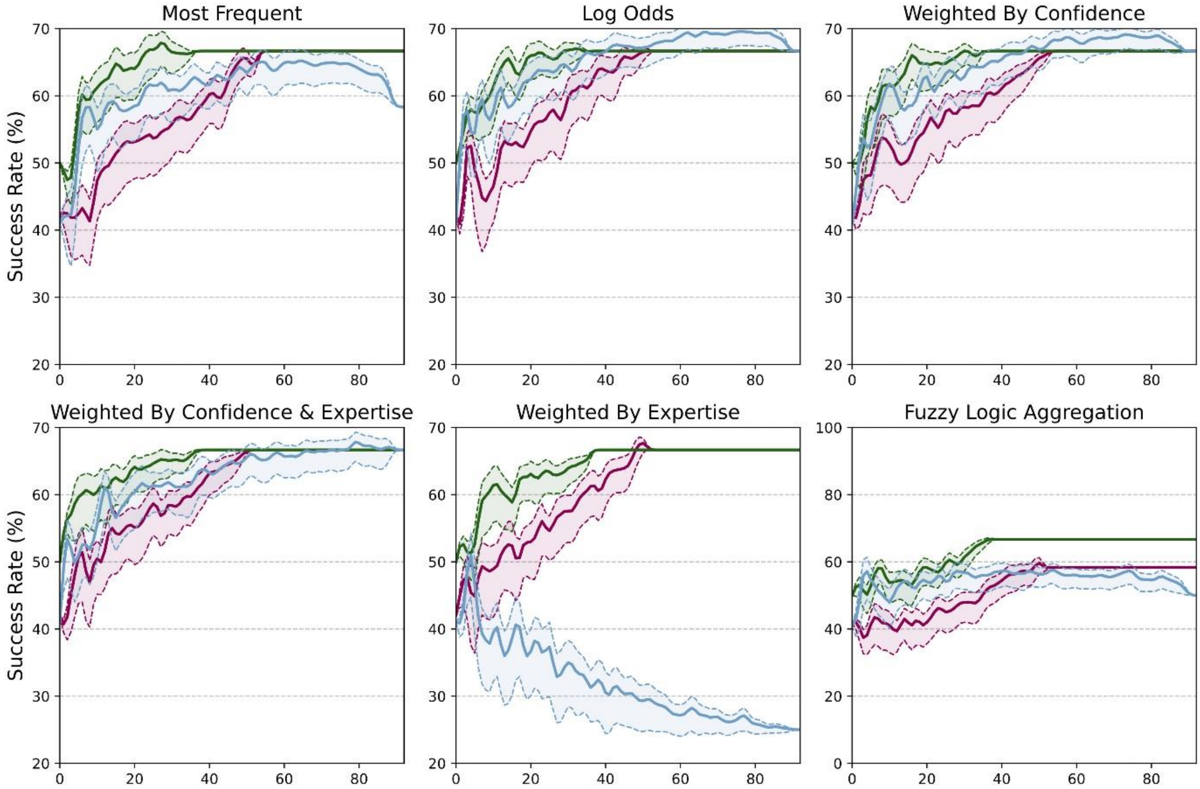


Figure S8: *Evolution of the collective success rate as a function of the number of participants in the population per aggregation method for the permeability endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2), expert (3-5), and all participants. Groups are colored as in Figure S6.

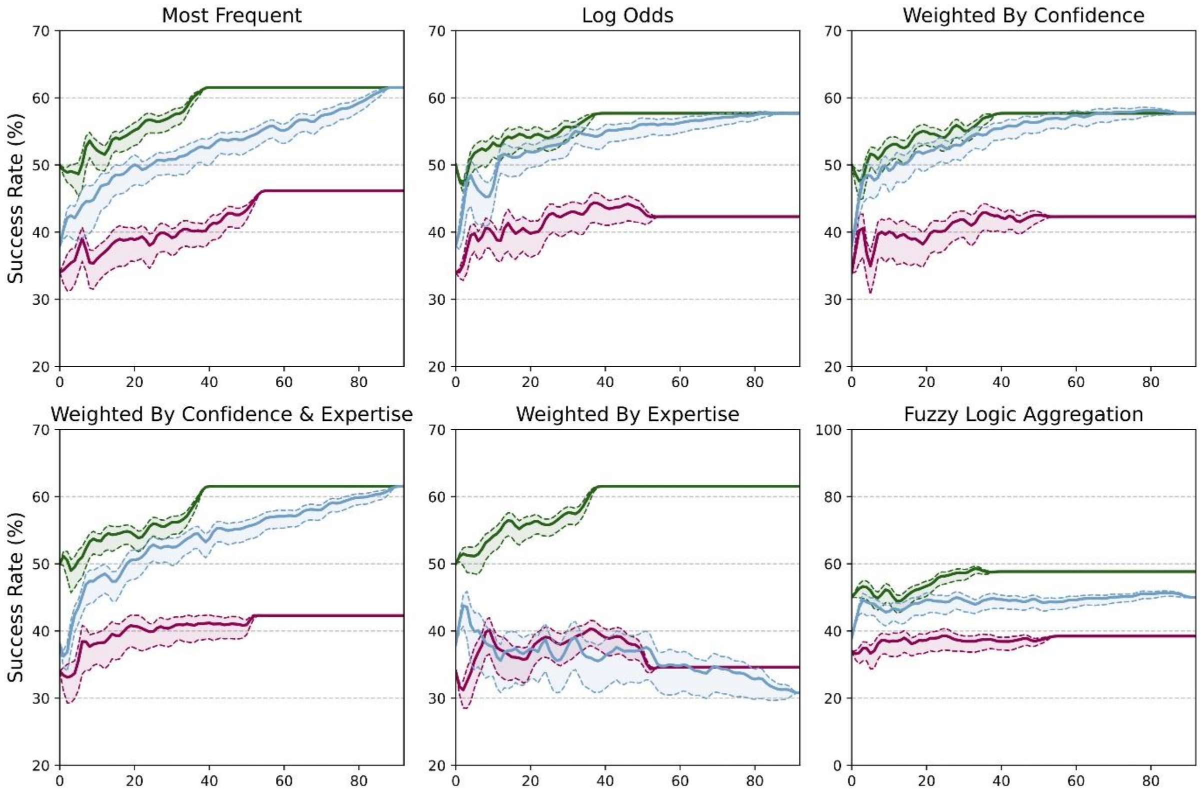


Figure S9: *Evolution of the collective success rate against the number of participants in the population per aggregation method for the solubility endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2), expert (3-5), and all participants. Groups are colored as in Figure S6.

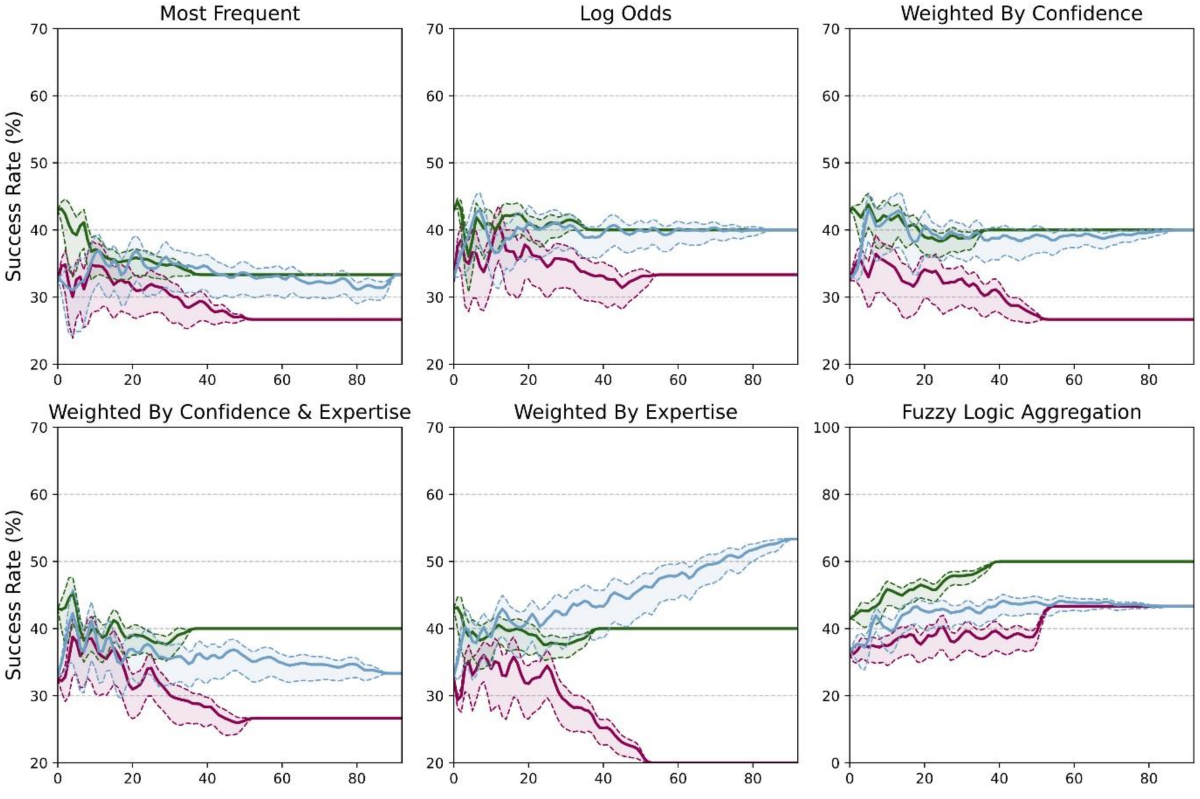


Figure S10: *Evolution of the collective success rate against the number of participants in the population per aggregation method for the hERG endpoint.* The collective answer is denoted per expertise group, from non-expert (1-2), expert (3-5), and all participants. Groups are colored as in Figure S6.

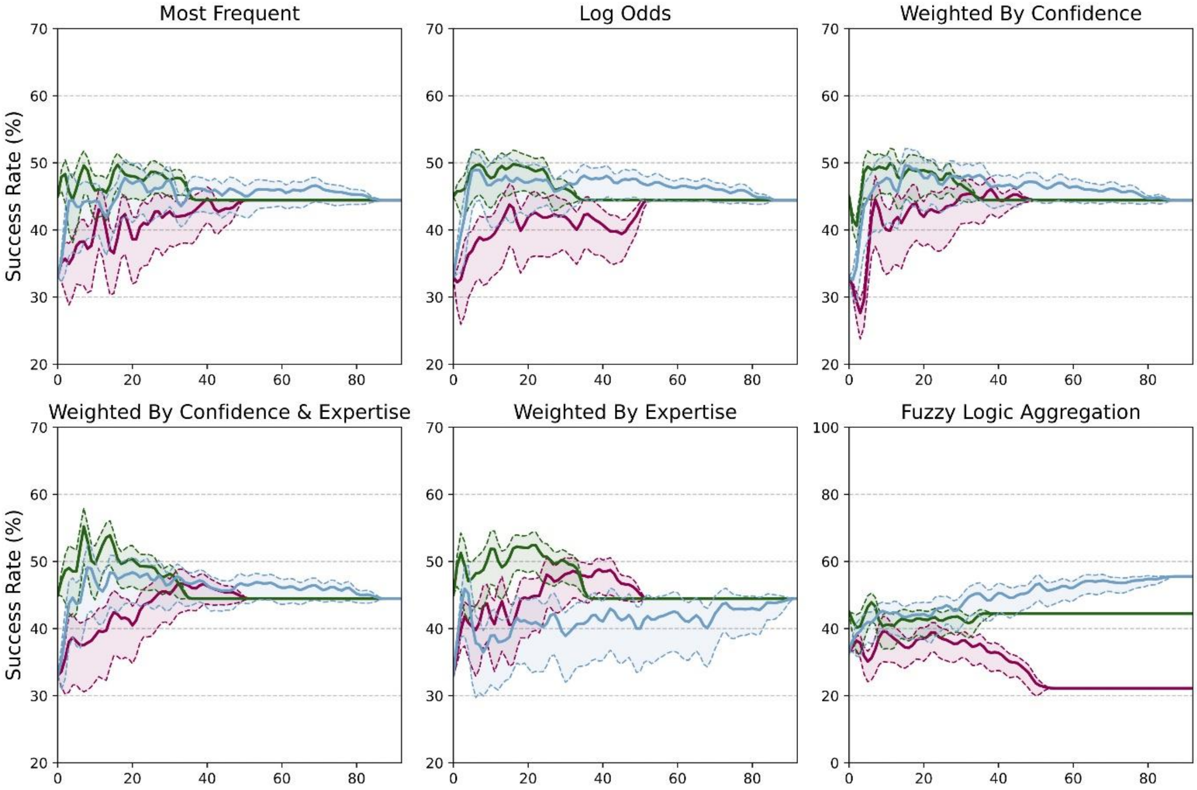


Figure S11: *Evolution of the collective success rate against the number of participants in the population per aggregation method for the LogD endpoint.* The collective answer is denoted per expertise group, from non-expert (1-2), expert (3-5), and all participants. Groups are colored as in Figure S6.

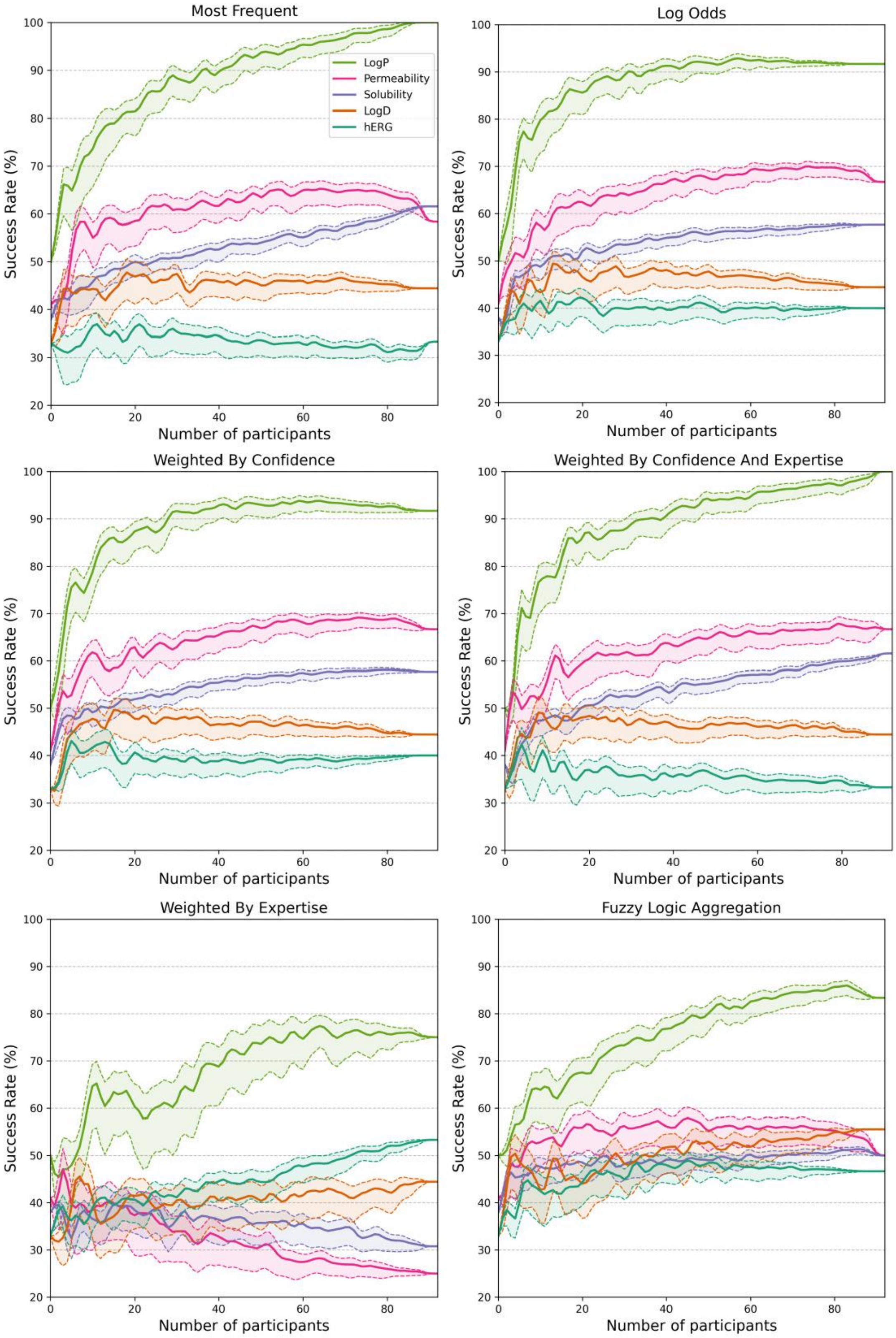


Figure S12: *Evolution of the collective success rate as a function of the number of participants in the population per endpoint and per aggregation method.*

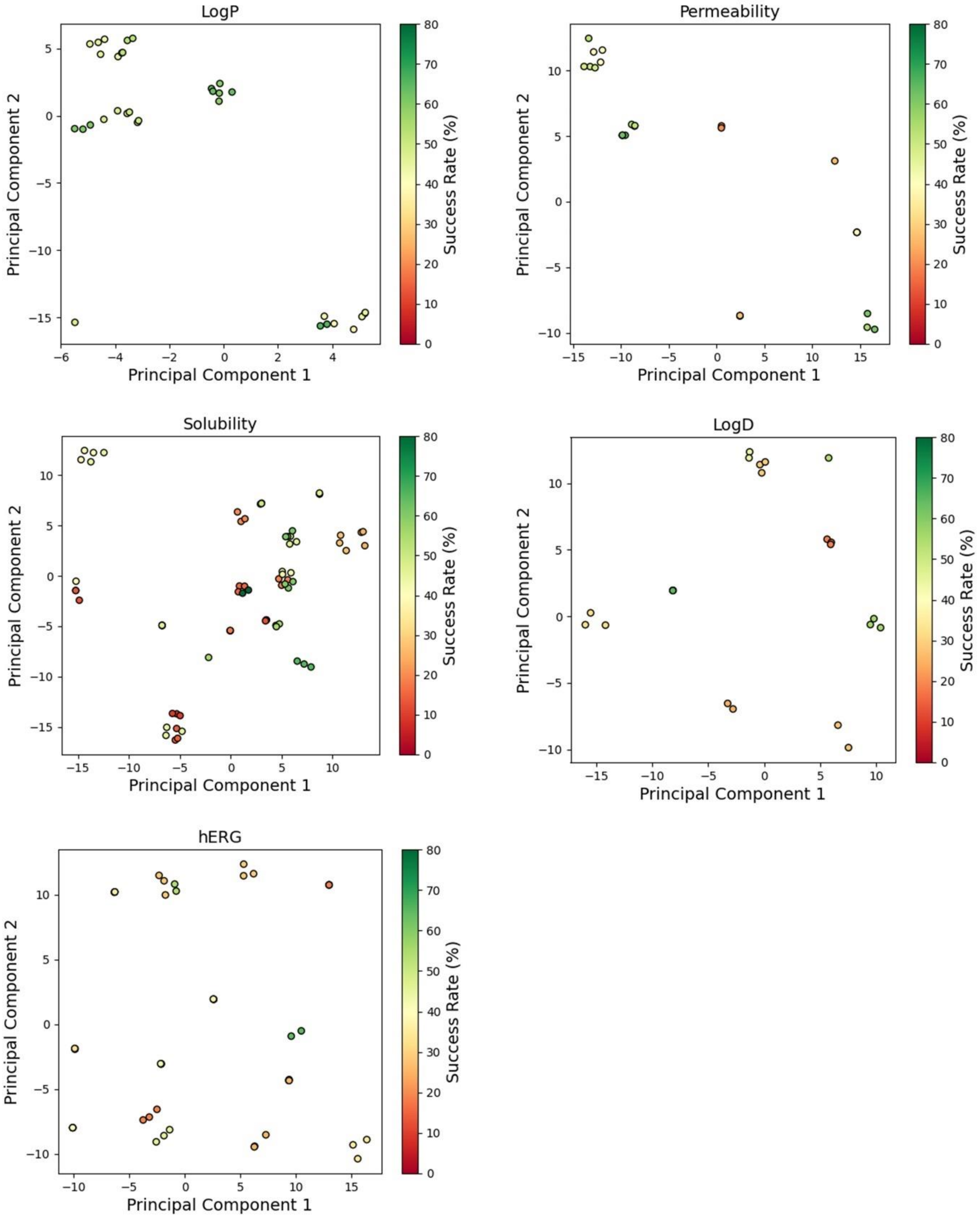


Figure S13: *t-SNE map of the collective intelligence chemical space per endpoint*. Each point represents a unique compound colored by the success rate of the related question.

Table S1: *Performance of the Graph Neural Networks on the public internal test set on ADMET endpoints.*

Endpoints	**Number of unique compounds**	**R^2**	**RMSE**	**MAE**
LogP	10,668	0.93	0.47	0.33
Permeability	1,259	0.55	0.56	0.42
Solubility	5,012	0.48	0.68	0.52
LogD	5,347	0.84	0.60	0.44
hERG	8,050	0.56	0.61	0.43

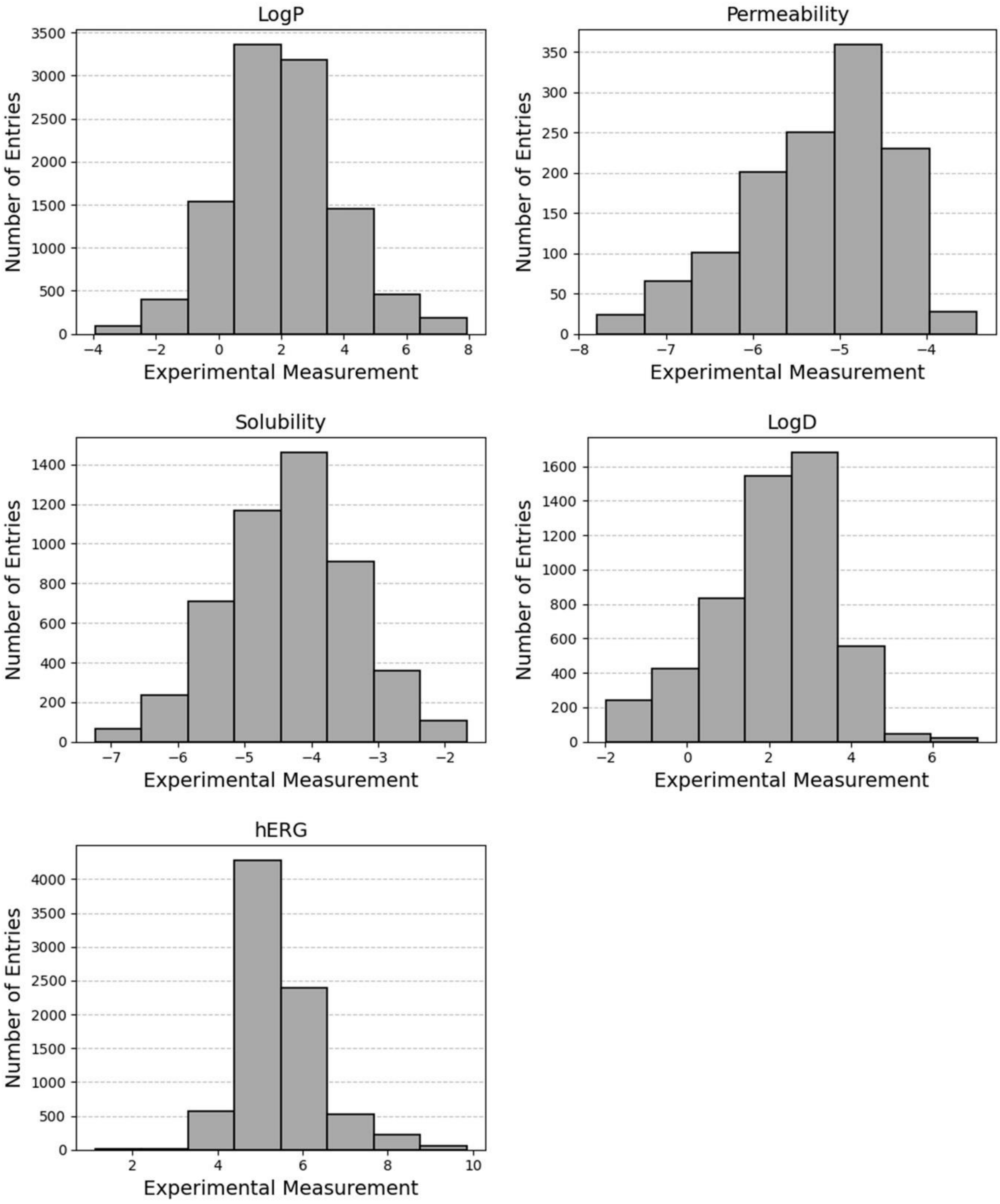


Figure S14: *Distribution of experimental measurements from public data used for modelling purposes.* The present endpoints are expressed in $\log_{10}(C_{octanol}/C_{water})$ for LogP, $\log_{10}$(cm/s) for permeability, $\log_{10}(C_{buffer})$ for solubility, $\log_{10}(C_{octanol}/C_{buffer})$ for LogD, and pIC50 for hERG inhibition.

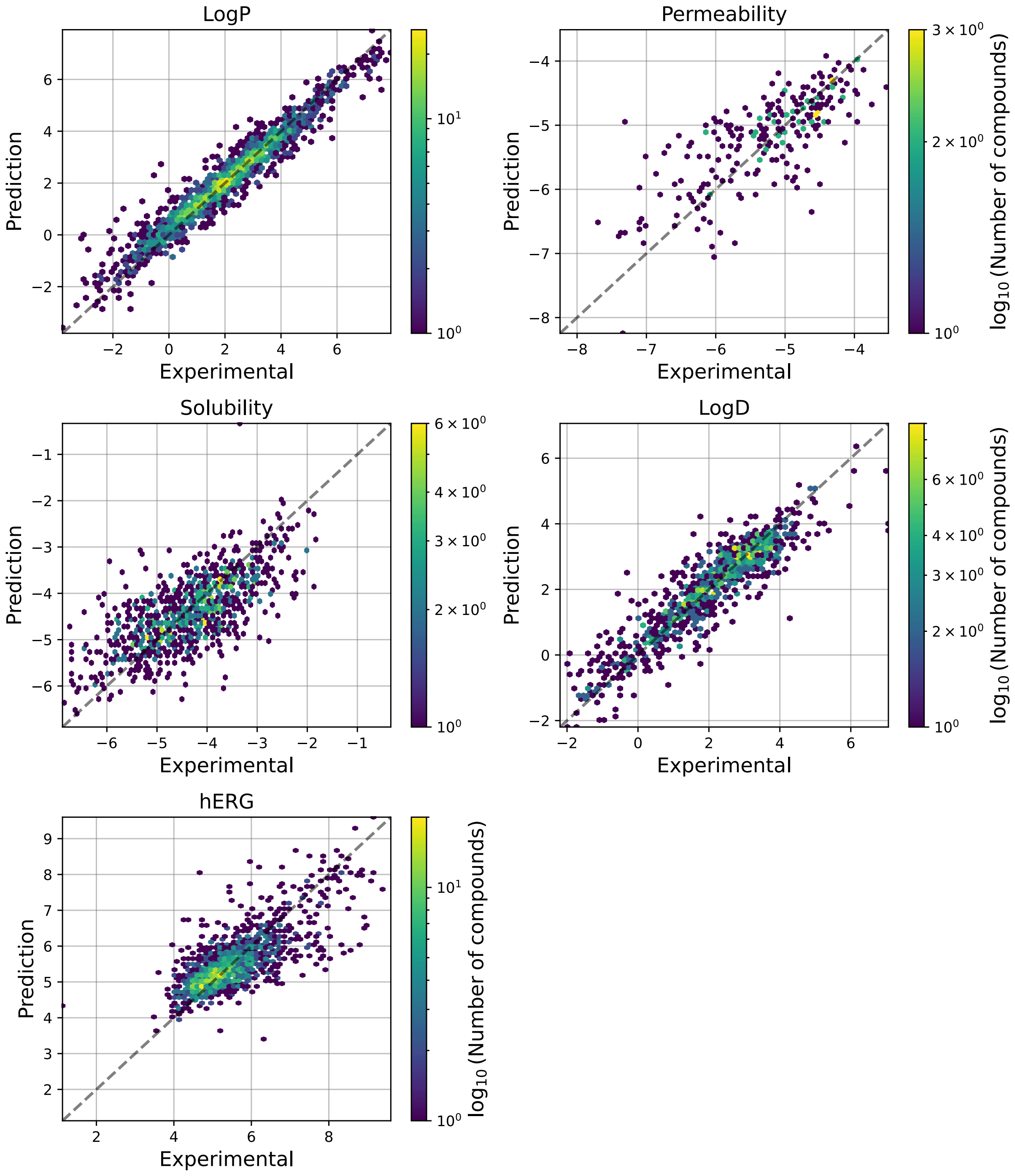


Figure S15: *Correlation between experimental and predicted value per endpoint.* The coloration depicts the density of compounds as the base-10 logarithm of the number of unique compounds. Answer success and failure count per ADMET endpoint. Answers are grouped per source such as human, GNN (predictive model), GNN & Human, and all.

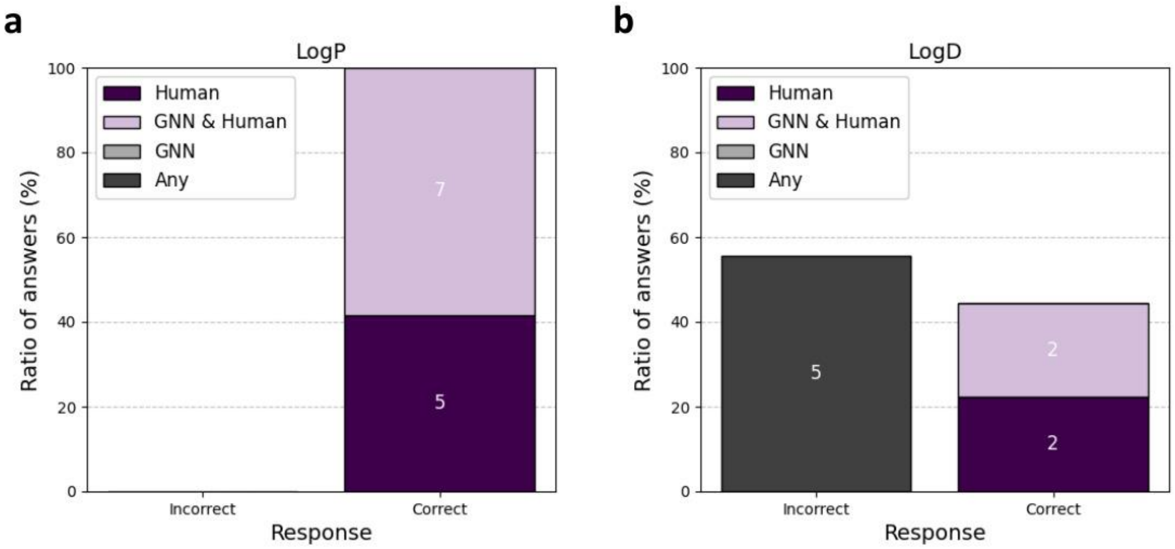


Figure S16: *Answer success and failure ratio (y-axis) and count (number in boxes) for a) logP and for b) logD.* Answers are grouped per source, *i.e.*, human, GNN (predictive model), GNN & Human, and all.