Learning chemical intuition from humans in the loop

Oh-Hyeon Choung[1], Riccardo Vianello[1], Marwin Segler[2], Nikolaus Stiefl[1,*], and José Jiménez-Luna[2,*]

[1]Novartis Institutes for Biomedical Research, 4002 Basel, Switzerland
[2]Microsoft Research Cambridge, CB1 2FB Cambridge, United Kingdom
[*]Correspondence: nikolaus.stiefl@novartis.com, jjimenezluna@microsoft.com

Abstract

The lead optimization process in drug discovery campaigns is an arduous endeavour where the input of many medicinal chemists is weighed in order to reach a desired molecular property profile. Building the expertise to successfully drive such projects collaboratively is a very time-consuming process that typically spans many years within a chemist's career. In this work we aim to replicate this process by applying artificial intelligence learning-to-rank techniques on feedback that was obtained from 35 chemists at Novartis over the course of several months. We exemplify the usefulness of the learned proxies in routine tasks such as compound prioritization, motif rationalization, and biased *de novo* drug design. Annotated response data is provided, and developed models and code made available through a permissive open-source license.

1 Introduction

Drug discovery is a complex, multi-step process that operates at the interface between many chemical and biological sub-disciplines. In many stages of the pipeline, and specifically during lead optimization, medicinal chemists — wet-lab or computational — play a central role, as they are routinely tasked with identifying which compounds to synthesize and evaluate over subsequent rounds of optimization.[1] In order to do this, medicinal chemists often review data that includes compound properties such as activity, ADMET,[2] or target structural information, among many others. Therefore, for a campaign to be successful it needs not only rely on the quality of the generated experimental data, but ultimately also on the robustness and soundness of the decisions made by the medicinal chemistry team working on it.[3]

During their professional careers, medicinal chemists build an expertise that enables them to make their decisions (*e.g.*, compound prioritization) more efficiently.[4] That is, they develop an "intuition" on the factors relevant for a compound to be successful on following iterations of the early drug discovery process. While attempts have been previously made to formalize such knowledge with rule-based approaches (*e.g.*, structural alerts), or simple cheminformatics desirability scores (*e.g.*, drug-likeness), capturing the subtleties and intricacies involved in the ranking ability of chemists remains a fundamental challenge. With that motivation in mind, in this work we investigate whether part of this knowledge can be distilled into machine learning models. Such models can potentially then be deployed as an aid in during the decision-making process in lead optimization or other parts of the drug discovery pipeline, similar to other recommendation systems already reported in the industry.[5–7]

Since medicinal chemistry is currently mostly a human endeavour, it is also inevitably prone to subjective biases.[8] Several studies[9,10] have evaluated to what degree medicinal chemists tend to agree on their own and the decisions made by their colleagues. Most tasks explored in these works included presenting chemists with a list of compounds to filter over several rounds, in order to evaluate whether their choices overlapped with those of their peers, and if they were self-consistent with their own prior selections. These studies reported overall a weak agreement between and within each chemist - the disparity in these results being associated to several psychological factors, such as loss aversion.[11] Another study,[12] closer in nature to what we present in this work, evaluated whether a small group of chemists could rate compounds according to properties such as drug-likeness and synthetic accessibility via the use of a Likert-type scale,[13] to then train a classical machine learning model on these responses. Fair to moderate correlations were found between the scores assigned by the chemists, but the reported study design could have been prone to the *anchoring* psychological

effect, in which decisions are affected by subject- and situation-specific reference values.[11] A recent work with a similar experimental setup was also described in the context of the design of porous organic cages.[14]

In this study we set to overcome those limitations by adopting a strategy that is well-known in the context of multiplayer games. We cast the goal of ranking a set of molecules as a preference learning problem and show that individual preferences can be captured via pairwise comparisons with a simple neural network architecture. 35 (wet-lab, computational, and analytical) chemists at Novartis participated in the study, with over 5000 annotations collected over several rounds driven by an active learning approach. We show that the learned implicit scoring functions capture aspects of chemistry currently not covered by other *in silico* chemoinformatics metrics and rule sets, some of them derived from highly-optimized internal annotations over years of cumulative *know-how*. We furthermore exemplify their applicability in the context of hit-to-lead compound prioritization and biased *de novo* machine-learning drug design. We also show that the proposed learned scoring function can better capture the concept of drug-likeness more accurately than another widely used metric (QED). We furthermore rationalize the learned chemical preferences by means of fragment analyses on a large public compound database. Finally, so as to facilitate reproducibility and foster additional research on this topic, a software package (MolSkill), containing production-ready models and anonymized response data, is made available through a permissive license in an accompanying code repository.

2 Materials

We organize this section by first describing the universe of participants in the study, as well as providing motivation on question design. We then provide details on the overall evolution and stages of the presented study, as several preliminary rounds were carried out to justify the main body of this work. Finally we then describe the different datasets used and their associated cleaning procedure.

2.1 User composition and question design

A total of 35 medicinal chemists from different sites at Novartis participated in the presented study. These included chemists from different geographical sites, at different levels of seniority/expertise, and from either a medicinal, organic, analytical, or computational chemistry background.

In regards to the question posed, and in the belief that chemists develop an inherent sense of what constitutes a desirable compound over their careers, we set out to present a fairly simple, and intentionally-ambiguous prompt: "Which of these two compounds do you prefer?". We asked chemists to imagine an early virtual screening campaign setting (accounting for simple aspects such as oral availability and small molecular profile, but no other modalities such as covalency or bifunctional compounds) where they needed to decide which compound to follow up between two. The question was designed so that participants did not spend a significant amount of time evaluating each presented pair of compounds, while being generic enough so that one of the compounds could be discarded according to a non-defined "gut feeling" chemical preference. This could include drug-likeness, synthetic accessibility or other criteria inherent to the pair of the compounds presented in each choice. We note that the question choice can be seen as an oversimplification of the problem, and that in other drug discovery scenarios, additional details on the presented prompt would be needed for clarification. Especially in real-life setups, these details would typically include aspects like existing ADMET or activity data, or bespoke predictive models for those endpoints.

2.2 Evolution of the study

Over the course of the presented study, several rounds were conducted. Two preliminary analysis rounds consisting of 220 molecule pairs, with feedback requested from 9 and 14 chemists at Novartis, respectively, were carried out. Specifically, we mainly focused on measuring:

- To what degree the choices made by one chemist agree with those made by their peers (*i.e.*, inter-rater agreement). This was evaluated with 200 different compound pairs. Intuitively this a direct measure of whether there is a signal to be learned by a machine-learning model.
- Whether chemists choices are self-consistent (*i.e.*, intra-rater agreement). In order to do so, we included an additional redundant 20 compound pairs, albeit in a random order and position on the screen.

Additionally, we also studied whether there was a bias towards choosing a compound depending on its position (left/right) during annotation. After the first initial preliminary round was completed, we had received qualitative feedback from the chemists on some of the presented pairs. Specifically some criticism was expressed in regards to some pairs being inherently hard, as both compounds contained clearly problematic motifs (*e.g.*, "plague *vs.* cholera"

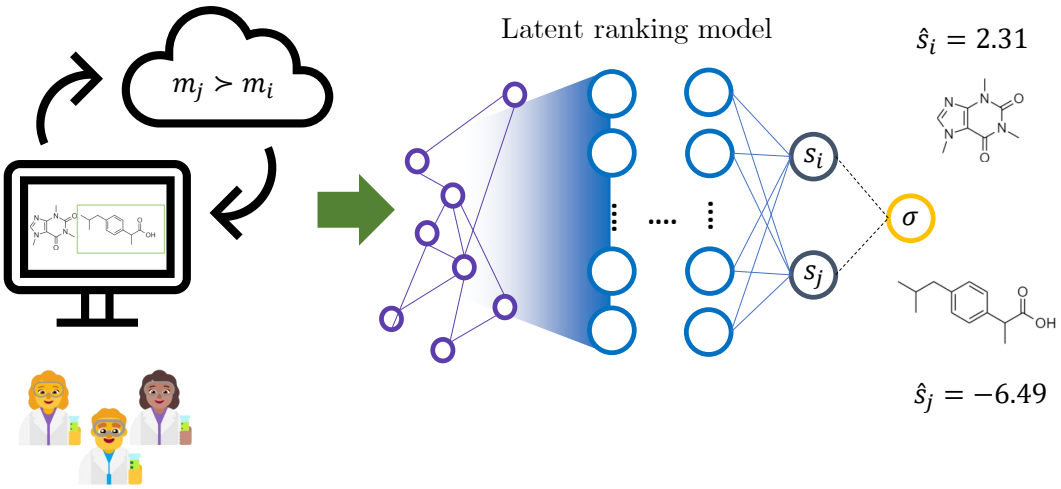

Figure 1: Overall schematic of the main idea behind the study. Molecules are treated as players playing competitive games against each other. The probability of one winning over the other is provided by feedback as supplied by chemists. For this, the chemists are asked to select one or the other depending on a pre-specified question prompt on a web application. An implicit score model is then learned based on this feedback, which can later be used for downstream cheminformatics tasks.

pairs where both compounds featured known toxicophores). These were then removed according to the procedure detailed in Section 2.3. A second round with identical number of pairs was subsequently carried out. Note that in the first and second preliminary rounds, all chemists were handed out the same pairs (*i.e.*, we performed inter-rater repetitions), so as to adequately evaluate the points presented above. After both preliminary rounds yielded satisfactory results (see Section 4.1), we set out for a production run where we obtained over 5000 responses over the course of several months. Furthermore, since a reasonable degree of agreement between the chemists in the preliminary rounds was observed, we forwent the pair repetition requirement in the production runs and considered all participating chemists as a single labeling oracle.

2.3 Data retrieval, cleaning, and pair generation

For all purposes of the study, we use compounds extracted from the publicly-available ChEMBL database[15] (version 30). Specifically, all compounds considered in this study come from a pool where the following filters were applied: their molecular weight was between 200 and 1000 g mol^{-1}, their drug likeness (QED)[16] between 0.2 and 0.9 and up to 2 rule-of-five violations.[17] Additionally, all retrieved compounds were checked so that they could successfully be read by the RDKit package,[18] and subsequently standardized, which included removal of salts, tautomer normalization,[19] and atom neutralization via O'Boyle's nocharge code.[20] For the second preliminary study round and subsequent production rounds (see Section 2.2), the NIBR substructure filters were also applied,[21] and compounds with more than 10 rotatable bonds or 3 fused rings were removed, which resulted in a final pool of $1,831,052$ molecules.

For the two preliminary study rounds, and for the first round of the production stage, the compounds present in the initial pool were grouped in 1000 clusters via the k-means algorithm, as implemented in scikit-learn,[22] and using binary extended-connectivity fingerprints as molecular features. Pairs were then selected by ensuring that their associated clusters were not repeated within the same batch of questions.

3 Methods

We begin by describing the user study design adopted here, and then follow by providing details on the neural network methodology applied to ranking compounds based on implicit user feedback. We finally describe the active learning strategy applied over the different rounds of the study and briefly showcase the developed platform for data collection.

3.1 Psychometric study setup

We considered a user study where interviewees were presented with pairs of choices (*i.e.*, compounds) to select from. There were several reasons to consider a pairwise experimental design in contrast to simpler alternatives such as obtaining direct feedback on individual samples. One of such advantages is that

there is plenty of evidence from psychometric studies and decision theory suggesting that humans find it inherently hard to sort items according to their preferences,[23] whereas making binary decisions is a task that is in general easier.[24–26] Additionally, it also avoids user or situation-specific baseline biases: humans are known to start labeling from an anchor value that is then adjusted towards a final decision in situations of uncertainty or stress. This has been demonstrated to be a major issue in other user studies.[27–31]

3.2 Learning to rank

Our setting resembles that of preference learning by pairwise comparisons.[32] One naïve approach to tackle the challenges raised by the proposed pairwise design is to try and induce a utility function based on how many times a compound has been preferred over others (or its proportion), and then frame this problem as a regular supervised regression task. The main disadvantage of this procedure, however, is that it requires the same compound to be present in several comparisons in order to accurately estimate a preference, which severely limits how much chemical space we can explore given a finite amount of time provided by the volunteer chemists. Instead, we take inspiration from the ELO skill-based systems that were popularized by the rating schemas for zero-sum games such as chess or backgammon,[33] or more recently by the TrueSkill algorithm[34,35] as used by the Xbox Live multiplayer videogame service. In the original setting, the difference in ratings between two players served as a function of the probability of one player winning over the other. In our case, we consider the presented molecules to the chemists as the "players" participating in our game, the main goal being to rank them[36] (Figure 1).

Mathematically, given a (possibly incomplete) set of molecules $m_1, m_2, \ldots, m_n \in \mathcal{M}$, and training data consisting of k pairs of examples with binary preference relations of the type $m_i \succ m_j$ (meaning that m_i was preferred over m_j in a specific match), our task is to infer a total ordering over all molecules in $\mathcal{M}$. Furthermore, such pairs do not need to specify a complete ranking of the training data or be consistent (*i.e.*, satisfy transitivity). In order to do so, we consider a function $s : \mathcal{M} \to \mathbb{R}$, where we assume that each molecule can be parameterized by a latent score that can be learned by a sufficiently-expressive model.[37] Once this function has been approximated, it can be then used to impose a complete order over already seen or new molecules. Denoting by $\delta_{ij} := \hat{s}(m_i) - \hat{s}(m_j)$ the learned latent score difference between molecules m_i and m_j, we estimate $\hat{p}\left(m_i \succ m_j\right) := \sigma\left(\delta_{ij}\right)$, where σ is a sigmoid function. To learn s, we then simply use standard stochastic gradient descent and minimize a binary cross-entropy loss between the probability estimates and the preference values in the training data. Since this loss is a function of the learned $\boldsymbol{\delta}$ values only, to ensure identifiability of the scores $\boldsymbol{s}$, and to guarantee that these are centered around the real origin, we use a regularization term $\mathcal{L}_{\text{reg}}(\hat{\boldsymbol{s}}; \lambda) := \lambda \left\|\hat{\boldsymbol{s}}\right\|^2$, where $\|\cdot\|$ is the Euclidean norm and λ is a user-defined hyperparameter. Empirically, we found that setting small values $\lambda \simeq 10^{-3}$ is enough to encourage the desired score behaviour for our use case.

We chose to parameterize s as a standard feed-forward neural network that uses 2048-bit count-based extended connectivity fingerprints[38] and a list of two-dimensional descriptors computed via RDKit as molecular input features. We train all models using the Adam[39] optimizer with an initial learning rate of 3×10^{-4}. Additional molecular featurization and architectural details are available in the accompanying code repository to this study.

3.3 Active learning

To achieve the set goal of 5000 user responses in the study, and to ensure we covered sufficient chemical space, we considered a simple batched active learning approach.[40,41] Specifically, every 1000 responses we randomly sampled a large number of pairs from the initial pool of compounds. These pairs were then ranked according to their uncertainty, as estimated by the variance of their predicted δ_{ij} values using the the Monte Carlo dropout method[42] with a fixed rate of 0.2 and 100 predicted samples. Additionally, to ensure that comparisons were not drawn between too many compounds belonging to similar regions of chemical space, we used the clustering strategy defined in Section 2.3 and allowed up to one comparison between any two clusters in each batch.

3.4 Platform deployment

A platform for questionnaire delivery was internally developed at Novartis. Users were asked to select between pairs of compounds presented upon a predefined question. The front-end was developed using an intuitive ReactJS[a] web GUI that could be operated either via a computer or a touchscreen device. A screenshot of the deployed interface is shown on Figure 2. Special care was taken to ensure that the same pair was not presented to different users. Results were internally stored in a remote PostgreSQL database[43] instance through a custom REST API developed with FastAPI[b]. The database was then peri-

[a] `reactjs.org`
[b] `fastapi.tiangolo.com`

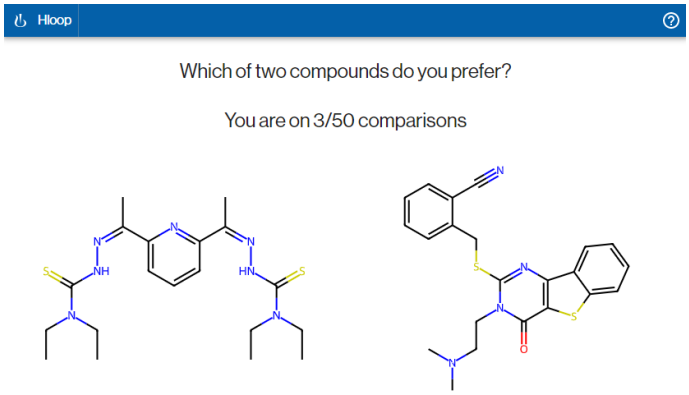

Figure 2: Screenshot of the web interface used for data collection during the study. Chemists were asked to select which of two compounds they preferred according to a prespecified question presented at the top of the page.

odically exported to perform model training and run analyses.

Table 1: Intra-rater agreement, as measured by the percentage of times chemists agreed with their previous choice on a pair and by the Cohen's κ coefficient. Left-right bias measured as the percentage of times a rater chose the compound presented on one side of the screen. Abbreviations: R1/R2: First/second preliminary round of the study

	Intra-rater Ag. (%)		Intra-rater Ag. (Cohen's κ)		Left-right bias (%)	
Chemist Id.	R1	R2	R1	R2	R1	R2
1	100.0	100.0	1***	1***	48.2	47.7
2	92.1	84.2	0.68***	0.37*	47.2	54.5
3	86.8	78.9	0.49*	0.16	48.6	55.5
4	79.8	84.2	0.27	0.35*	54.6	48.2
5	85.1	92.1	0.37*	0.69***	47.2	48.6
6	89.5	-	0.55**	-	47.2	-
7	84.2	92.1	0.33	0.65***	48.6	47.7
8	94.7	92.1	0.79***	0.69***	46.8	51.8
9	95.6	89.5	0.89***	0.58***	50.9	48.2
10	-	81.6	-	0.28	-	52.7
11	-	92.1	-	0.69***	-	53.2
12	-	92.1	-	0.69***	-	56.8
13	-	92.1	-	0.69***	-	50.9
14	-	89.5	-	0.58**	-	51.8
15	-	94.7	-	0.79***	-	54.1

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

4 Results

We first focus on the evaluation of the results provided by the first preliminary rounds for the study, which ultimately led us to pursue the subsequent production-level runs. This is followed by a quantitative evaluation of predictive model performance over the different rounds considered in the study. We then proceed to explore several areas where we believe the proposed scoring function can be practical. We study the relationship of the learned scoring function to other common *in silico* metrics in chemoinformatics and evaluate whether it can distinguish between chemical sets of different nature. We further investigate whether more precise learned chemical preferences can be rationalized via means of a fragment analysis and, finally, exemplify the usage of the proposed scoring function in biased molecular generation.

4.1 Preliminary analysis rounds

Results for the two preliminary rounds are summarized in Table 1. As a measure for inter-rater agreement, we consider the Fleiss' κ_F coefficient[44] among the responses provided by the chemists in both preliminary rounds. We measured $\kappa_{F1} = 0.4$ and $\kappa_{F2} = 0.32$ for the first and second round, respectively, and concluded that there was a moderate agreement between the preferences expressed by the chemists. One likely reason for the only moderate agreement is the fact, that especially in cases where there is no clear-cut preference, decisions were be driven by prior personal experiences. Still, these results suggested that there was a pattern to be learned by the responses to the posed question. Using the redundant pairs present in both preliminary rounds, we also evaluated intra-rater agreement between the chemists using the Cohen's κ_C coefficient. With $\kappa_{C1} = 0.6$ and $\kappa_{C2} = 0.59$ for the first and second preliminary round, respectively, we conclude that that in most cases, all chemists displayed a fair degree of response consistency. Additionally, no specific positional bias on screen was detected for any of the preliminary participants, with preferences reasonably close to the expected random 50% baseline. Additional two-by-two inter-rater agreement coefficients are presented in Figure S1, from which we draw similar conclusions.

Overall, the results on the preliminary rounds suggested that there was indeed a signal to be learned from the opinions expressed by the chemists that had participated in the study up to that point. These findings convinced us to extend the study and continue with the subsequently presented, larger production-level runs.

4.2 Predictive pair preference performance

In order to evaluate whether the trained model successfully learned the preferences expressed by the chemists, we iteratively measured its predictive performance via the area under the receiver-operating characteristic (AUROC) curve under different scenarios (Figure 3). Specifically, we kept the data from

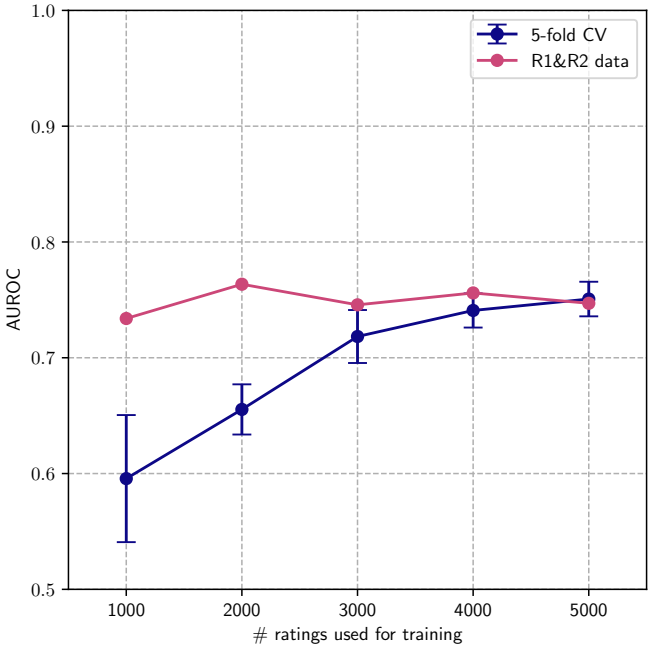

Figure 3: Predictive performance of the proposed latent score ranking model when evaluating which compounds are preferred within each pair. Results presented at different train set sizes corresponding to the associated active learning batches considered during the study.

the preliminary rounds as external sets for validation that are not used for model training or uncertainty quantification during the active learning rounds. Additionally, we also evaluated model performance via 5-fold cross-validation after each labeled batch of 1000 samples. From the cross-validation results, a steady pair classification performance improvement can be observed as more data becomes available, starting from 0.6 and surpassing 0.74 AUROC values at the 1000 and 5000 available pairs thresholds, respectively. Interestingly, cross-validation results did not display reaching a performance plateau even when evaluated at the last available batch of responses, hinting that performance could further be improved if more data had been collected. Model results stayed relatively stable around the 0.75 AUROC value when evaluated on the preliminary round data, which could be explained by the limited amount of pairs available in these sets. Overall, these results suggest that the model is able to correctly learn preferences as expressed by medicinal chemists in the current experimental setup. For completeness, we also evaluated to what degree different common molecular representations had an impact on model overall performance (Figure S2, Table S1).

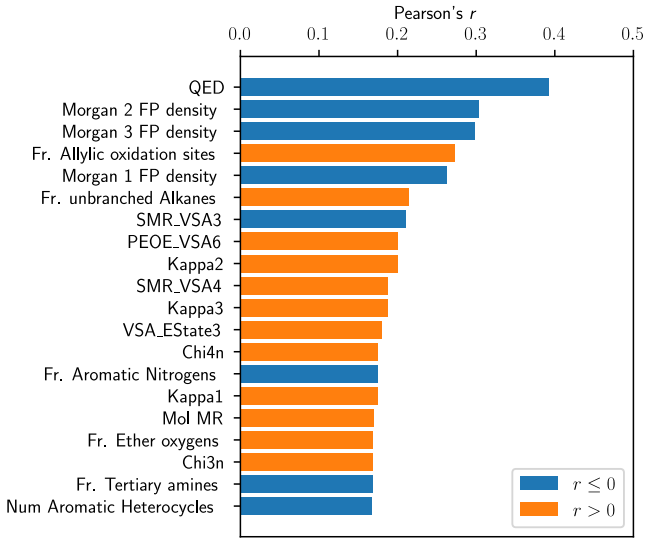

Figure 4: Correlation coefficients between several *in silico* descriptors computed via RDKit and learned compound scores in the training set. Results shown for the 20 most correlated *in silico* metrics (in absolute value).

4.3 Relationship to other *in silico* metrics

One of the main assumptions of the main question presented to the participants in this study is that, over the course of their careers, medicinal chemists develop an expertise that is hardly quantifiable by other existing *in silico* metrics. In order to evaluate whether such is the case, we measure to what degree the learned compound scores correlate with other ligand-based properties that are commonly-used during optimization (*e.g.*, drug-likeness, topological surface area, number of saturated rings). All properties considered were computed with the RDKit software package. A summary of the highest correlated properties (on an absolute scale) in the training data is presented in Figure 4. With Pearson correlation coefficients overall not surpassing the $r = 0.4$ threshold, we conclude that the learned scores are in fact providing a perspective on molecules that is orthogonal to what can be currently computed with other cheminformatics software routines. Among the most correlated properties we can find: drug-likeness,[16] fingerprint density, the fraction of allylic oxidation sites, atomic contributions to the van der Waals surface area,[45] or the Hall-Kier kappa value.[46] For completeness, an extensive list of all of the properties computed as well as their correlations to the learned scores is also provided in Figure S3.

4.4 Discriminating between chemical sets

As a way of quantitatively evaluating whether the learned scores can be used to deprioritize compounds

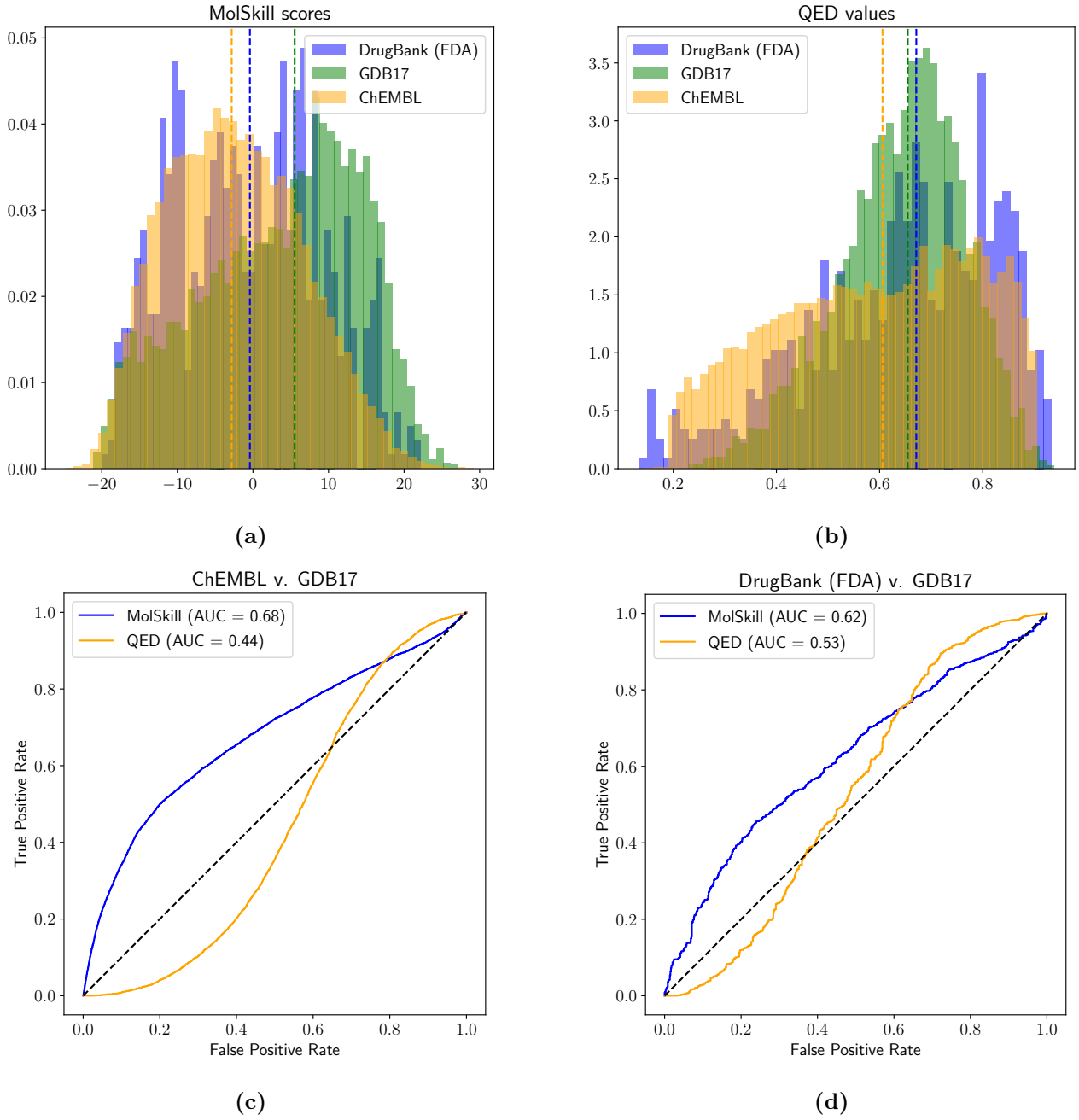

Figure 5: (a-b) Distribution of MolSkill scores and QED values over three different molecular sets: ChEMBL, a set of FDA-approved drugs as made available by DrugBank, and a random sample of the combinatorially-generated GDB17 dataset. (c-d) ROC AUC curves for both MolSkill scores and QED values when tasked to discriminate between molecules from either ChEMBL or FDA-approved drugs from GDB17-extracted molecules

that could be seen as undesirable, we consider an approach similar to that one reported in the original QED study.[16] Specifically, we scored different sets of molecules: a random subset of 30,000 ChEMBL compounds present in the original pool for this study, a set of FDA-approved drugs as made available by the DrugBank[47] database, and a random subset of 10,000 compounds extracted from the GDB17 database.[48] Furthermore, we used the latter GDB17 compounds as a control, since it was originally generated in a combinatorial fashion and should in practice also contain molecules that do not exhibit drug-like properties. Furthermore, to ensure that the molecules considered in these analyses did not fall out of the applicability domain of the trained model, we made sure to apply the same filtering procedures as those detailed in Section 2.3, resulting on 732, and 4889 analyzed molecules for the FDA-approved drugs and GDB sets, respectively. As a baseline method to compare the learned scores with, we considered the standard QED implementation as available on the RDKit package. On Figure 5a, it can be observed that the distribution of learned scores is clearly well separated between sets better representing drug-like space (in other words

those more apealing to medicinal chemists, *i.e.*, Drugbank FDA-approved drugs and ChEMBL) against the GDB17 set. QED scores (Figure 5b) on the other hand, struggle at making such separation between the three sets. While an three-way ANOVA test was performed and the null hypothesis of equal mean values was rejected for both methods with virtually zero p-values ($F_{\text{MolSkill}} = 945.69$, $F_{\text{QED}} = 178$), receiver operating characteristic curves to distinguish the drug-like sets against GDB17 showed that only the proposed learned scores were predictive enough in practice for this task (Figures 5c, d).

4.5 Exploring fragment preference

As means for interpretability towards what structural information the proposed model has learned over the course of the study, in this section we aim to disentangle whether it has developed a preference for specific molecular motifs. In order to do so, we make use of the BRICS algorithm,[49] as implemented in the RDKit software, and compute all available leaf fragments and associated model scores for each molecule present in the training set. Since the fragments contained an "attachment" atom type not seen during training, fragments were scored according to the average scores of the compounds they were substructures of in the training set. Additionally, to avoid biases related to uncommon motifs or unexplored areas of chemical space, only fragments appearing a minimum of 5 times in the training set were considered in this analysis. A small selection of the highest and lowest ranked fragments is presented in Figure 6. Among the worst ranked fragments we can observe undesirable groups such as phenols, free acids, ketones, thioureas, allyls, long alkyl chains, naphtyls, cumarines, Hantzsch esters, quaternary amines, sugar-like structures or highly-substituted rings. On the other hand, among the best-ranked groups we can find many commonly-used medicinal chemistry motifs such as pyrazine, pyrimidine, sulfones, imidazoles, oxadiazoles, phenyls, or bicyclic heteroaromatics. Qualitatively, this experiment suggest that the proposed score has learned patterns that are in line with groups present in existing drug-like molecules. The full set of fragments, the frequency of their occurrence, as well as their associated score is provided in the accompanying code repository to this work.

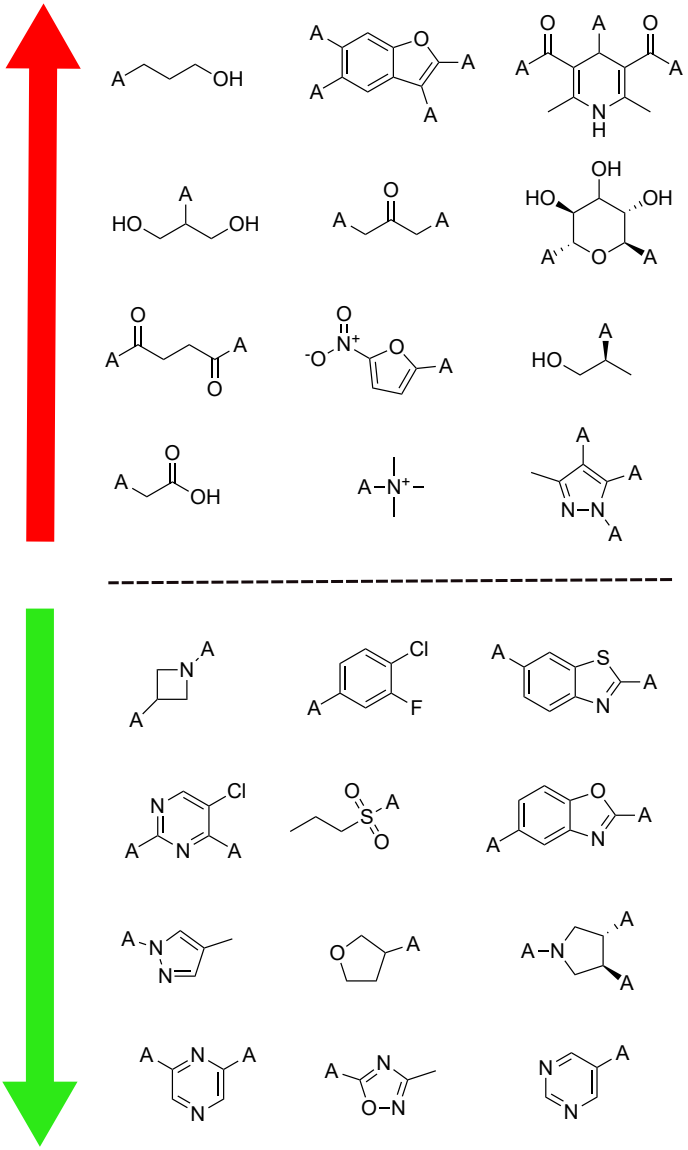

Figure 6: Some fragment examples evaluated by the learned scoring function. Fragments representative at each end of the score distribution (lower is better)

4.6 Biased molecular design

As a way of exemplifying how the implicitly-learned scoring function may be applied in a realistic setting, in this section we use it to bias a generative model towards favorable regions of chemical space. We make use of the GuacaMol baselines[50] package and implemented a submodule with the proposed scoring function trained on all available rating data. We then chose the pretrained SMILES-based LSTM generative model and the hill-climbing optimization strategy[51] to generate 500 molecules both maximizing and minimizing the learned scoring function. Some generated molecule examples are presented in Figure 7. Visually inspecting some of the examples maximized by the scoring function, we can appreciate that the model is assigning high scores (*i.e.*, "unattractive") to compounds that feature long flexible chains, atypical groups such as phosphates or azides, conjugated double bonds and reactive pieces, and overall higher number of carboxylates and alcohols, among many other non-drug like properties. On the other hand, minimizing the learned scoring function results in a reasonable mix of aromatic rings and aliphatic sp^3 carbons, reasonably-sized fragments as well as several typical groups featured in drug-like molecules. From

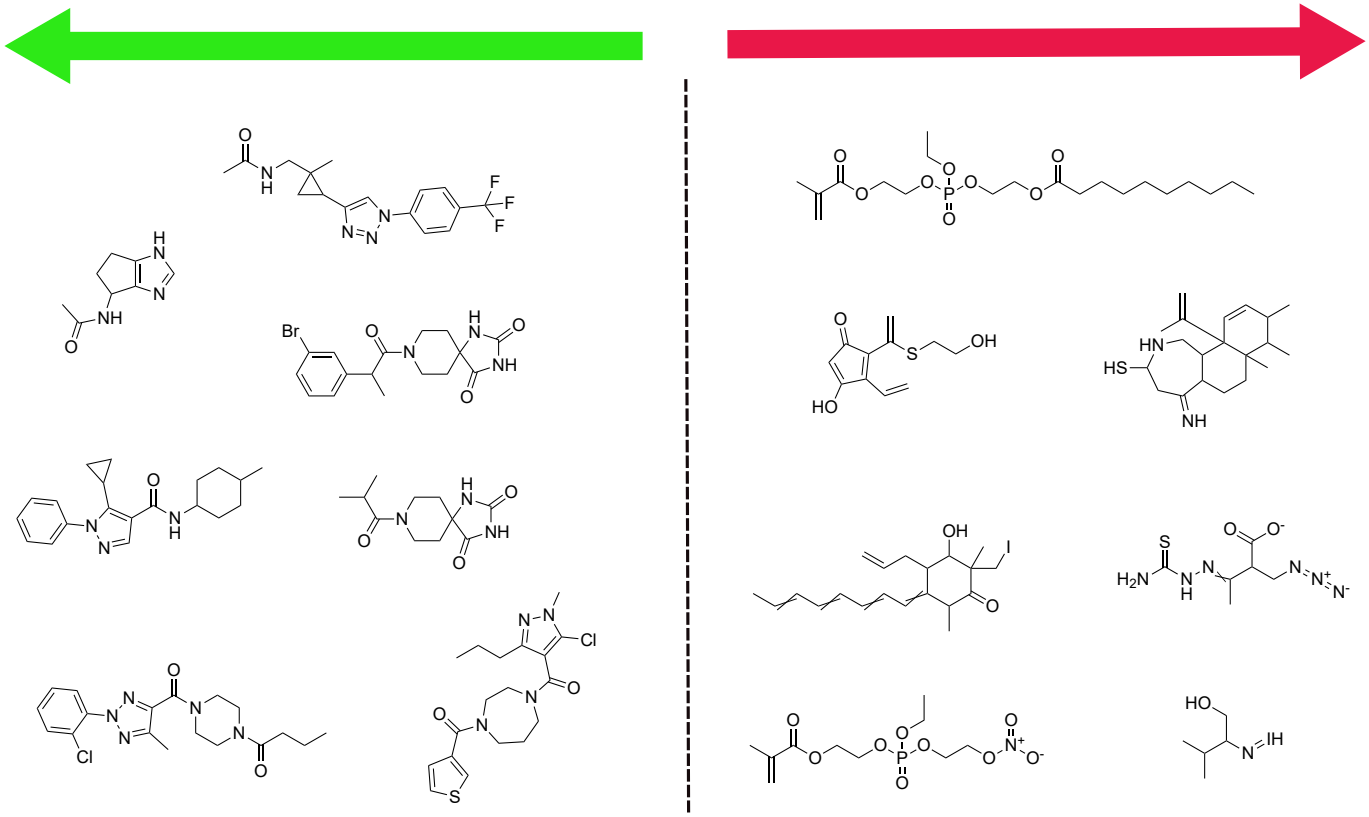

Figure 7: Some molecular examples prioritized by the proposed implicit scoring function when paired with a generative model. Results presented for both maximization and minimization of the learned score (lower is better).

these qualitative analyses we conclude that the scoring function has successfully captured a reasonable degree of chemical intuition.

One caveat that we had experimentally observed during molecular generation is that it was useful to constrain or stop optimization of the scoring function once it had reached values close to the limits to the empirical distribution of learned scores ($|\hat{s}| \approx 9$ using the reported regularization strategy during training in our sets). Not doing so resulted in a certain degree of quirkiness and invalid molecules, which we attribute to the generative algorithm overexploiting the scoring function on regions of chemical space that it had not previously observed during training. Additional details on the generative model and optimization hyperparameters are made available in the accompanying code repository to this work.

4.7 Qualitative score assessments on ChEMBL

While the quality of the generated molecules indicate a high relevance of the proposed scoring function for *de novo* drug design, we furthermore qualitatively evaluated its usefulness to filter out undesirable compounds. This was studied especially in the light of existing rule-based approaches, such as the NIBR filters,[21] which are routinely used to deprioritize and flag problematic compounds before consideration. Ideally, our goal was to rationalize compound features not necessarily captured by such methods currently, but that are at the same time considered as undesirable by medicinal chemists. Towards this goal, we manually reviewed molecules from the initial pool, which had been already filtered with simple properties as well as with the aforementioned rules, and then visually inspected ones that were assigned a high score by the proposed function. Figure 8 shows four of such compounds. While there are some features that could be described as unattractive and can be captured with a generic SMARTS pattern (*e.g.*, the terminal alkene in compound *b*, or the aromatic nitro group in compound *c*), the overall "unattractiveness" seems to be driven by more general properties. Based on our subjective opinion, among others these seem to include compound complexity (*c* and partially *a*), a mix of flexibility and feature-richness (*b* and *d*), or the distribution of features (*b*). While theoretically possible, defining such rules explicitly is a difficult task and is unlikely to capture all undesirable cases.

5 Discussion

In this work, we have proposed and described the development of a machine-learned scoring function

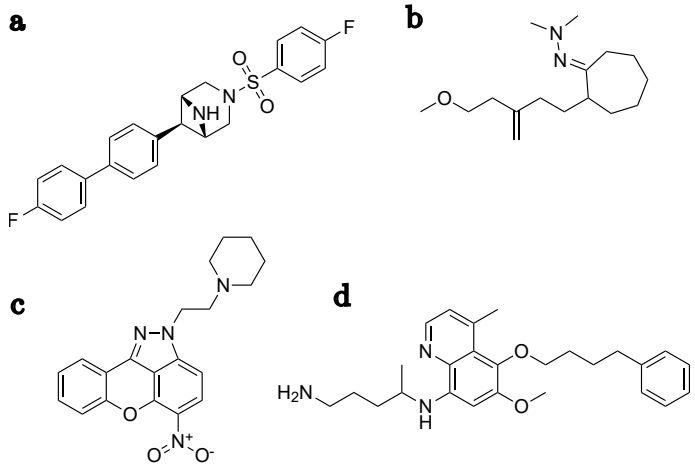

Figure 8: Example compounds not flagged by the NIBR filters but effectively deprioritized by the learned score (9 or higher).

of human preference in the context of early drug discovery campaigns. We have done so by adapting the well-known framework of player ratings to a pairwise learning-to-rank experimental design between molecules. In order to do so, we have internally deployed a large user study at Novartis, where the expertise of 35 medicinal chemists was taken into account. To the best of our knowledge, this is the first study of its kind, where we show that such expertise can be successfully learned by a latent score machine-learning model. Such scores have been shown to be providing additional or orthogonal information to what can be obtained by other common *in silico* ligand-based properties or substructure-based fragment definitions. We furthermore exemplify the utility of such modeling approach in several routine cheminformatics tasks, such as the deprioritization of compounds currently not flagged by well-known rule-based approaches, or biased biased molecular design via a generative ML model. We furthermore rationalized and motivated what the model has learned by means of a fragment analysis on a large set of compounds and show that it outperforms a popular quantitative measure of drug-likeness at distinguishing chemical sets of different nature.

We see the utility of the proposed model to go beyond what is proposed in the current study. Specifically, we believe that there is potential to extend the discussed setup for other observables in drug discovery that are inherently quantifiable but expensive to obtain experimentally (*e.g.*, compound stability calculations). Additionally, we believe it could provide insights into unexplored regions of chemical space currently ignored when applying simpler mnemonics such as Lipinski's rule of five.[52,53] With that in mind, we believe that "soft" versions of some popular rule-based filters can be learned by artificially generating training pairs alongside a similar architecture as the one proposed. Such models could potentially overcome the main limitation of having to pre-filter compounds before inference so as to avoid out-of-distribution risks. Another main limitation of the study relates to the simplicity of the question asked during data collection, which was left intentionally vague to capture chemical intuition on a timely manner. Ultimately, whether the proposed function can be successfully used as an aid to experts at any stage in the drug discovery process, and especially in a prospective fashion, remains a topic of further study.

Data & code availability

All data associated with this study, including survey responses and production-ready trained models are available via a MIT-licensed repository `github.com/microsoft/molskill`. A conda package is also provided for integration convenience within downstream cheminformatics tasks. Neural network models were trained using the PyTorch automatic differentiation library (version 1.11).

Conflict of interest

None to declare.

Author contributions

O.C.: platform deployment, analysis, manuscript writing; R.V.: platform deployment; M.S.: conceptualization, manuscript writing; J.J-L.: supervision, conceptualization, experimental design, open sourcing, manuscript writing; N.S.: supervision, experimental design, manuscript writing.

Acknowledgements

We are grateful for the feedback (and criticism!) received from all the participant chemists at Novartis, without whom this study would not have been possible. We also thank the entire GenChem team at both Novartis and Microsoft Research for helpful discussions on this work.

References

[1] Veale, C. G. Into the fray! A beginner's guide to medicinal chemistry. *ChemMedChem* **16**, 1199–1225 (2021).

[2] Van De Waterbeemd, H. & Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nature Reviews Drug Discovery* **2**, 192–204 (2003).

[3] Gomez, L. Decision making in medicinal chemistry: The power of our intuition (2018).

[4] Cheshire, D. R. How well do medicinal chemists learn from experience? *Drug Discovery Today* **16**, 817–821 (2011).

[5] Rohall, S. L. *et al.* An artificial intelligence approach to proactively inspire drug discovery with recommendations. *Journal of Medicinal Chemistry* **63**, 8824–8834 (2020).

[6] Boström, J., Falk, N. & Tyrchan, C. Exploiting personalized information for reagent selection in drug design. *Drug Discovery Today* **16**, 181–187 (2011).

[7] Vidler, L. R. & Baumgartner, M. P. Creating a virtual assistant for medicinal chemistry. *ACS Medicinal Chemistry Letters* **10**, 1051–1055 (2019).

[8] Leeson, P. D., Davis, A. M. & Steele, J. Drug-like properties: Guiding principles for design–or chemical prejudice? *Drug Discovery Today: Technologies* **1**, 189–195 (2004).

[9] Kutchukian, P. S. *et al.* Inside the mind of a medicinal chemist: The role of human bias in compound prioritization during drug discovery. *PloS ONE* **7**, e48476 (2012).

[10] Lajiness, M. S., Maggiora, G. M. & Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *Journal of Medicinal Chemistry* **47**, 4891–4896 (2004).

[11] Kahneman, D. & Tversky, A. Choices, values, and frames. *American Psychologist* **39**, 341 (1984).

[12] Takaoka, Y. *et al.* Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *Journal of Chemical Information and Computer Sciences* **43**, 1269–1275 (2003).

[13] Likert, R. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[14] Bennett, S. *et al.* Materials precursor score: Modeling chemists' intuition for the synthetic accessibility of porous organic cage precursors. *Journal of Chemical Information and Modeling* **61**, 4342–4356 (2021).

[15] Gaulton, A. *et al.* Chembl: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107 (2012).

[16] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry* **4**, 90–98 (2012).

[17] Lipinski, C. A. Lead-and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technologies* **1**, 337–341 (2004).

[18] RDKit: Open-source cheminformatics. `http://www.rdkit.org`. [Online; accessed 11-April-2013].

[19] Baker, C. M. *et al.* Tautomer standardization in chemical databases: Deriving business rules from quantum chemistry. *Journal of Chemical Information and Modeling* **60**, 3781–3791 (2020).

[20] O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 1–14 (2011).

[21] Schuffenhauer, A. *et al.* Evolution of Novartis' small molecule screening deck design. *Journal of Medicinal Chemistry* **63**, 14425–14447 (2020).

[22] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[23] Rosenman, R., Tennekoon, V. & Hill, L. G. Measuring bias in self-reported data. *International Journal of Behavioural & Healthcare Research* **2**, 320 (2011).

[24] Bradley, R. A. & Terry, M. E. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345 (1952).

[25] Roy, B. Classement et choix en présence de points de vue multiples. *Revue Française d'Informatique et de Recherche Opérationnelle* **2**, 57–75 (1968).

[26] Behzadian, M., Kazemzadeh, R. B., Albadvi, A. & Aghdasi, M. PROMETHEE: A comprehensive literature review on methodologies and applications. *European Journal of Operational Research* **200**, 198–215 (2010).

[27] Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* **185**, 1124–1131 (1974).

[28] Englich, B. & Soder, K. Moody experts—How mood and expertise influence judgmental anchoring. *Judgment and Decision Making* **4**, 41 (2009).

[29] Barbosa, S. D., Fayolle, A. & Smith, B. R. Biased and overconfident, unbiased but going for it: How framing and anchoring affect the decision to start a new venture. *Journal of Business Venturing* **34**, 528–557 (2019).

[30] McElroy, T. & Dowd, K. Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment and Decision Making* **2**, 48 (2007).

[31] Danziger, S., Levav, J. & Avnaim-Pesso, L. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* **108**, 6889–6892 (2011).

[32] Fürnkranz, J. & Hüllermeier, E. Preference learning and ranking by pairwise comparison. In *Preference Learning*, 65–82 (Springer, 2010).

[33] Elo, A. E. *The rating of chessplayers, past and present* (Arco Pub., 1978).

[34] Herbrich, R., Minka, T. & Graepel, T. Trueskill™: A Bayesian skill rating system. *Advances in Neural Information Processing Systems* **19** (2006).

[35] Minka, T., Cleven, R. & Zaykov, Y. Trueskill 2: An improved Bayesian skill rating system. *Technical Report* (2018).

[36] Chu, W. & Ghahramani, Z. Preference learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, 137–144 (2005).

[37] Burges, C. *et al.* Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 89–96 (2005).

[38] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754 (2010).

[39] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[40] Settles, B. Active learning literature survey (2009).

[41] Zhdanov, F. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954* (2019).

[42] Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059 (PMLR, 2016).

[43] Stonebraker, M., Rowe, L. A. & Hirohama, M. The implementation of POSTGRES. *IEEE Transactions on Knowledge and Data Engineering* **2**, 125–142 (1990).

[44] Fleiss, J. L. & Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* **33**, 613–619 (1973).

[45] Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences* **39**, 868–873 (1999).

[46] Kier, L. & Hall, L. The kappa indices for modeling molecular shape and flexibility. In *Topological Indices and Related Descriptors in QSAR and QSPAR*, 465–500 (CRC Press, 2000).

[47] Wishart, D. S. *et al.* Drugbank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

[48] Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling* **52**, 2864–2875 (2012).

[49] Degen, J., Wegscheid-Gerlach, C., Zaliani, A. & Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery* **3**, 1503–1507 (2008).

[50] Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling* **59**, 1096–1108 (2019).

[51] Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science* **4**, 120–131 (2018).

[52] Hartung, I. V., Huck, B. R. & Crespo, A. Rules were made to be broken. *Nature Reviews Chemistry* (2023).

[53] Leeson, P. D. & Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery* **6**, 881–890 (2007).