

Molecular deep learning at the edge of chemical space

Received: 1 August 2025

Accepted: 6 March 2026

Published online: 22 April 2026

 Check for updatesDerek van Tilborg^{1,2}, Luke Rossen¹ & Francesca Grisoni^{1,2}  

Molecular machine learning models often fail to generalize beyond the chemical space of their training data, limiting their ability to reliably perform predictions on structurally novel bioactive molecules. Here, to advance the ability of machine learning to go beyond the ‘edge’ of their training chemical space, we introduce a joint modelling approach that combines molecular property prediction with molecular reconstruction. This approach allows the introduction of unfamiliarity, a reconstruction-based metric that enables the estimation of model generalizability. Via a systematic analysis spanning more than 30 bioactivity datasets, we demonstrate that unfamiliarity not only effectively identifies out-of-distribution molecules but also serves as a reliable predictor of classifier performance. Even when faced with the presence of strong distribution shifts on large-scale molecular libraries, unfamiliarity yields robust and meaningful molecular insights that go unnoticed by traditional methods. Finally, we experimentally validate unfamiliarity-based molecule screening in the wet lab for two clinically relevant kinases, discovering seven compounds with low micromolar potency and limited similarity to training molecules. This demonstrates that unfamiliarity can extend the reach of machine learning beyond the edge of the charted chemical space, advancing the discovery of diverse and structurally novel molecules.

Molecular machine learning is rapidly gaining traction in early drug discovery^{1–5}. One key objective is identifying novel bioactive molecules (‘hits’) on one or more pharmacological targets⁶. In this context, finding structurally novel hit molecules is crucial for addressing unmet therapeutic needs^{7,8}, ensuring commercial viability⁹ and overcoming drug resistance^{10,11}. However, moving beyond the structural features of the training molecules (for example, to identify novel bioactive molecular cores) poses a substantial challenge for machine learning models, which often fail when applied to out-of-distribution (OOD) molecules^{12–14}. This is especially true for discrete data, such as molecules, which can quickly deviate from the data distribution learned during model training. This is further exacerbated by the scarcity of structurally diverse molecular data with high-quality experimental annotations^{12–15}, due to the costly and time-consuming nature of biochemical experiments. As a result, training sets typically contain only

hundreds of molecules, while libraries used for screening may contain billions of existing, but previously unseen, chemicals to be predicted¹⁶. Dealing with the resulting distribution shifts makes the discovery of structurally novel hit molecules with machine learning a herculean task. In this regard, quantifying how reliable predictions are beyond the ‘edge’ of the explored chemical space holds enormous promise.

Ensuring prediction reliability in prospective hit-screening campaigns has been an active topic of research^{17,18}. A well-established approach involves defining an applicability domain^{19,20}, which delimits the chemical space of reliable predictions, most often via a threshold on molecular similarity to the training data²¹ (Fig. 1a). However, this method does not incorporate the information learned by the model, and, due to its similarity-based definition, hampers the discovery of structurally novel molecules. Another widely used approach is based on uncertainty estimation^{22,23}, which leverages the model’s prediction

¹Institute for Complex Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands.

²Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Utrecht, the Netherlands. ✉ e-mail: f.grisoni@tue.nl

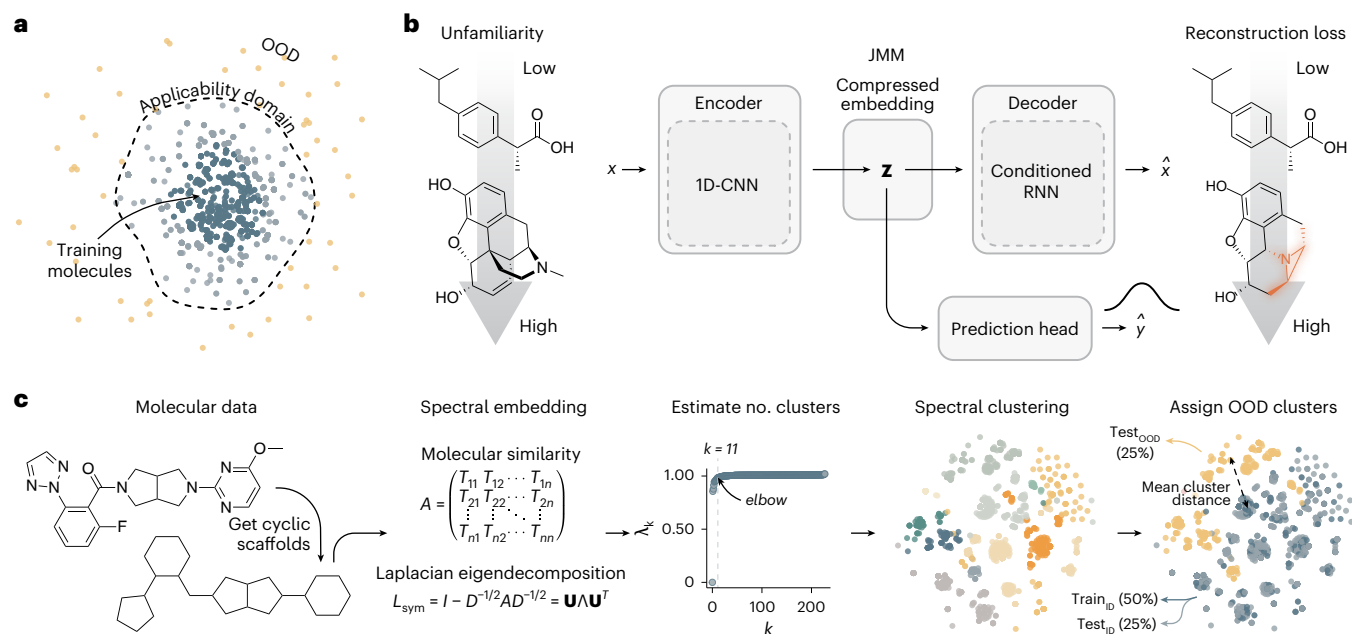


Fig. 1 | Estimating unfamiliarity of molecular data using joint modelling. **a**, Conceptual representation of the applicability domain. Molecules close to the training data in chemical space are within a model's applicability domain. Molecules outside of this boundary are considered OOD. **b**, The architecture of the JMM estimates how 'unfamiliar' a molecule is to the model through its

reconstruction loss. **c**, Inducing molecular distribution shifts by separating molecular data into in-distribution and OOD groups through spectral clustering. Results for the Orexin receptor 2 (OX2R) dataset are shown. CNN, convolutional neural network. RNN, recurrent neural network.

confidence, often through probabilistic modelling techniques. While uncertainty estimation allows the consideration of structurally novel molecules in principle, it may provide overconfident predictions when confronted with OOD samples^{14,24,25}. Hence, being able to make reliable predictions on OOD molecules remains one of the core challenges of molecular machine learning in drug discovery.

Here we offer a fresh perspective on how to better navigate the 'edge' of chemical space with deep learning, while accounting for prediction reliability on OOD molecules. To achieve this, we leverage recent advances in generative deep learning for de novo molecule design^{2,26–29}, in particular autoencoders^{27,30,31}. Autoencoders can be trained to encode molecular structures into a lower-dimensional latent space, and subsequently decode them back to their original form^{27,30}. In this work, through joint molecular modelling (Fig. 1b), we simultaneously train deep learning models to predict molecular properties (for example, bioactivity) and reconstruct the input molecule in a semi-supervised^{32,33} manner, that is, by learning from a combination of labelled and unlabelled molecular data.

Our joint learning approach breaks with the well-established application of leveraging a self-supervised learning task for generative chemistry²⁷ or for predictive performance improvement^{33–37}, by using reconstruction capabilities as a direct proxy for OOD estimation^{38–41}. Specifically, we hypothesize that poorly reconstructed molecules are less familiar to the model, indicating that they fall outside the distribution learned from the training data. Building on this hypothesis, we introduce a metric, termed unfamiliarity, which captures a model's reconstruction ability and is proposed to quantify how much a molecule deviates from the training distribution.

In this systematic study, spanning 33 experimentally labelled molecular datasets, we show not only that the introduced unfamiliarity metric is a robust indicator of molecular distribution shifts, but also that it strongly correlates with classifier performance. The capacity of unfamiliarity to identify structurally diverse and bioactive molecule is further validated in the wet lab, discovering several compounds with low micromolar activity on two kinase proteins.

Ultimately, the introduced concept of molecular unfamiliarity provides a principled approach to estimating model generalizability, even in the presence of molecular distribution shifts. Our approach offers a fresh perspective on estimating prediction reliability, complementing established concepts such as the applicability domain^{19,20} and uncertainty estimation^{22,23}—guiding the discovery of structurally novel molecules in a more precise and informed manner.

Results

In what follows, we will elaborate on our joint molecular model (JMM), introduce the unfamiliarity metric and demonstrate its ability to quantify molecular distribution shifts. Next, we relate molecular property prediction to distribution shifts and leverage this relationship to estimate prediction reliability. Finally, we apply the unfamiliarity metric to prioritize molecules in a virtual screening case study, followed by experimental validation in the wet lab.

Unfamiliarity and joint molecular modelling

The JMM (Fig. 1b) is based on a semi-supervised^{32,33} autoencoder, building on seminal work in molecular generative modelling²⁷. First, molecules are represented as Simplified Molecular Input Line Entry System (SMILES) strings⁴², which encode molecular topology and atom/bond types in a textual format. SMILES strings are encoded into a compressed latent vector (\mathbf{z}), using a one-dimensional convolutional neural network⁴³, which was shown to capture elements of bioactivity effectively^{27,44}. The \mathbf{z} vector is then decoded back into the input representation using a recurrent neural network with long short-term memory (LSTM)^{26,28,29}, in a self-supervised manner. This encoder-decoder was pretrained on ~1.2 M unlabelled molecular structures from ChEMBL⁴⁵, a dataset large enough to learn the 'grammar' of SMILES strings⁴⁶. The model was then finetuned using each labelled molecule, by passing the same latent representation (\mathbf{z}) to an approximate Bayesian classifier⁴⁷, which predicts molecular properties and simultaneously estimates prediction uncertainty (Fig. 1b). Reconstruction and property prediction were trained jointly (Methods;

equation (13)), to ensure that the shared latent space captures relevant information for both tasks.

The performance of molecular reconstruction was quantified via the reconstruction loss. This was computed as the total negative log-likelihood loss of all SMILES string tokens, normalized by SMILES token length:

$$\mathcal{L}_{\text{reconstruction}}(x) = -\frac{1}{|x|} \sum_{i \in x} \log p(t_i | t_{<i}), \quad (1)$$

where t_i is the i th non-padding element ('token') in the input SMILES string (x), and $|x|$ is the number of non-padding tokens in the sequence. Here, $p(t_i | t_{<i})$ denotes the probability assigned by the decoder to the next token t_i given all preceding tokens in the sequence. From the reconstruction loss, we obtain the unfamiliarity metric (\mathbb{U}), as follows:

$$\mathbb{U}(x) = \log \mathcal{L}_{\text{reconstruction}}(x). \quad (2)$$

The unfamiliarity metric depends on vocabulary size V and theoretically ranges from near $-\infty$ to $\log(\log(V))$. The lower the reconstruction loss, the lower the unfamiliarity, and vice versa. In our case, a model with $V = 35$ would have an upper \mathbb{U} limit of -1.27 under ideal circumstances with uniformly predicted token probabilities. However, in practice, this upper bound can be exceeded when a model assigns extremely low probabilities to certain tokens.

Detecting molecular distribution shifts

To investigate whether the unfamiliarity score reflects molecular distribution shifts, we collected 33 experimentally annotated datasets from literature^{48–50}, spanning various biological properties and sizes (Supplementary Table 1). These datasets were split into groups of in-distribution and OOD molecules. First, we performed spectral clustering (Methods) on molecular cyclic skeletons⁵¹ (core ring systems without exocyclic substituents), which ensured that structurally similar molecules were consistently grouped together (Fig. 1c). Based on cluster distances in each dataset, the most distant clusters (representing approximately 25% of all molecules) were used as an OOD test set (test_{OOD}). The remaining molecules were split into an in-distribution test set (test_{ID} , ~25% of total) and a training set (train_{ID} , ~50%; Supplementary Fig. 1).

To confirm that the molecules in the test_{OOD} set originate from different data distributions than the other two sets, we used three approaches to capture molecular similarity⁵²:

- (1) Similarity of extended connectivity fingerprints⁵³ (ECFPs), which capture the presence of atom-centred substructures. ECFPs were computed on molecular cyclic skeletons, and similarity was quantified via the Tanimoto coefficient⁵⁴.
- (2) Overlap of structural cores⁵⁵, computed as the maximal common substructure (MCS) fraction between molecular graphs (Methods).
- (3) Pharmacophore similarity, computed on Chemically Advanced Template Search (CATS) descriptors⁵⁶, using the cosine similarity.

Across all similarity metrics, molecules in the test_{OOD} set exhibited statistically significant differences from the molecules in the training set and test_{ID} , whereas no differences were found between these latter two (Fig. 2a–c, paired, two-sided, Wilcoxon signed-rank test, $\alpha = 0.05$).

To benchmark the effect of such a distribution shift on predictive performance, we trained three well-established⁴⁸ molecular property prediction models using molecular descriptors: a random forest (RF) and a multilayer perceptron (MLP) combined with ECFPs, as well as a RF model combined with CATS pharmacophore descriptors. As expected, the performance of these baselines consistently degraded on test_{OOD} compared with test_{ID} (Fig. 2d, paired, two-sided, Wilcoxon

signed-rank test, $P < 0.05$) as measured via balanced accuracy (Methods; equation (15)). Such performance degradation is a hallmark of OOD data⁵⁷.

Having established a baseline, we evaluated the classification performance of the proposed JMM (Fig. 1b) on each dataset. The JMM achieved a balanced accuracy of 0.75 ± 0.02 on test_{ID} molecules, which was slightly lower than the balanced accuracy of 0.78 ± 0.02 achieved by the ECFP-based model (Fig. 2d, $P = 3.7 \times 10^{-5}$, paired, two-sided, Wilcoxon signed-rank test). Such a small performance gap suggests limited practical differences, and it aligns with existing literature on the performance of SMILES-based versus ECFP-based models⁴⁴. Notably, training the JMM with or without the reconstructive decoder had no effect on the performance (test_{ID} : $P = 0.499$, test_{OOD} : $P = 0.594$; paired, two-sided, Wilcoxon signed-rank test). These results indicated that the performance of the JMM is in line with literature standards, despite the addition of the molecular reconstruction task. In other words, the decoder enables unfamiliarity estimation without imposing a performance penalty on the classifier.

Next, we aimed to verify the hypothesis that molecules well-represented in the training data (or test_{ID}) are reconstructed better by the JMM than 'unfamiliar' OOD molecules^{38–41}. To this end, we inferred $\mathbb{U}(x)$ on all labelled molecules in the respective test_{ID} and test_{OOD} sets for every trained JMM. As expected, test_{OOD} molecules received significantly higher unfamiliarity scores than test_{ID} molecules (two-sided Kolmogorov–Smirnov (KS) test, $P < 0.05$; Fig. 2e), as was especially visible on a dataset basis (Fig. 2f). Importantly, these differences in unfamiliarity were not driven by factors like SMILES string length or complexity, branching, molecular graph complexity, molecular weight or the number of functional groups (Supplementary Fig. 2). These results suggest that the model's ability to reconstruct molecules depends not on molecular complexity, but rather on its 'proximity' to the training data distribution.

For both molecules in test_{ID} and test_{OOD} , this was further confirmed by the direct relationship between a molecule's unfamiliarity and its distance to the training data, regardless of the data splits (Supplementary Fig. 3), for example, in terms of structural core overlap (Fig. 2g,h). Moreover, we found that, in general, there is a moderate-to-strong relationship between unfamiliarity and structural distance to the training data (Table 1). In this context, unfamiliarity correlates well to multiple and complementary similarity metrics, suggesting that it provides a generalizable, model-driven perspective on learned data distributions.

Unfamiliarity and bioactivity prediction

Because molecular reconstruction and molecular property prediction depend on the same learned latent representation, we tested how informative $\mathbb{U}(x)$ is for assessing the predictive capabilities of the model. We compared our approach with other well-established measures of reliability, namely:

- (1) Similarity to training set molecules (data-driven), measured as the average pharmacophore similarity, cyclic skeleton similarity, or molecular core overlap.
- (2) Embedding distance (model-driven), defined as the average Mahalanobis distance⁵⁸ of a molecule's embedding \mathbf{z} (Fig. 1b) to the learned embeddings of the training set.
- (3) Prediction uncertainty (model-driven), based on approximate Bayesian modelling of the classifier (Methods; equation (12)).

We found that all tested reliability metrics are indicative of a model's predictive performance, considering balanced accuracy, hit rate and precision (Supplementary Fig. 4). Prediction uncertainty and unfamiliarity have a moderate to strong correlation to model performance (Table 2). In other words, when uncertainty and/or unfamiliarity are high (on a dataset level), models make more erroneous bioactivity predictions, whereas molecules for which these metrics

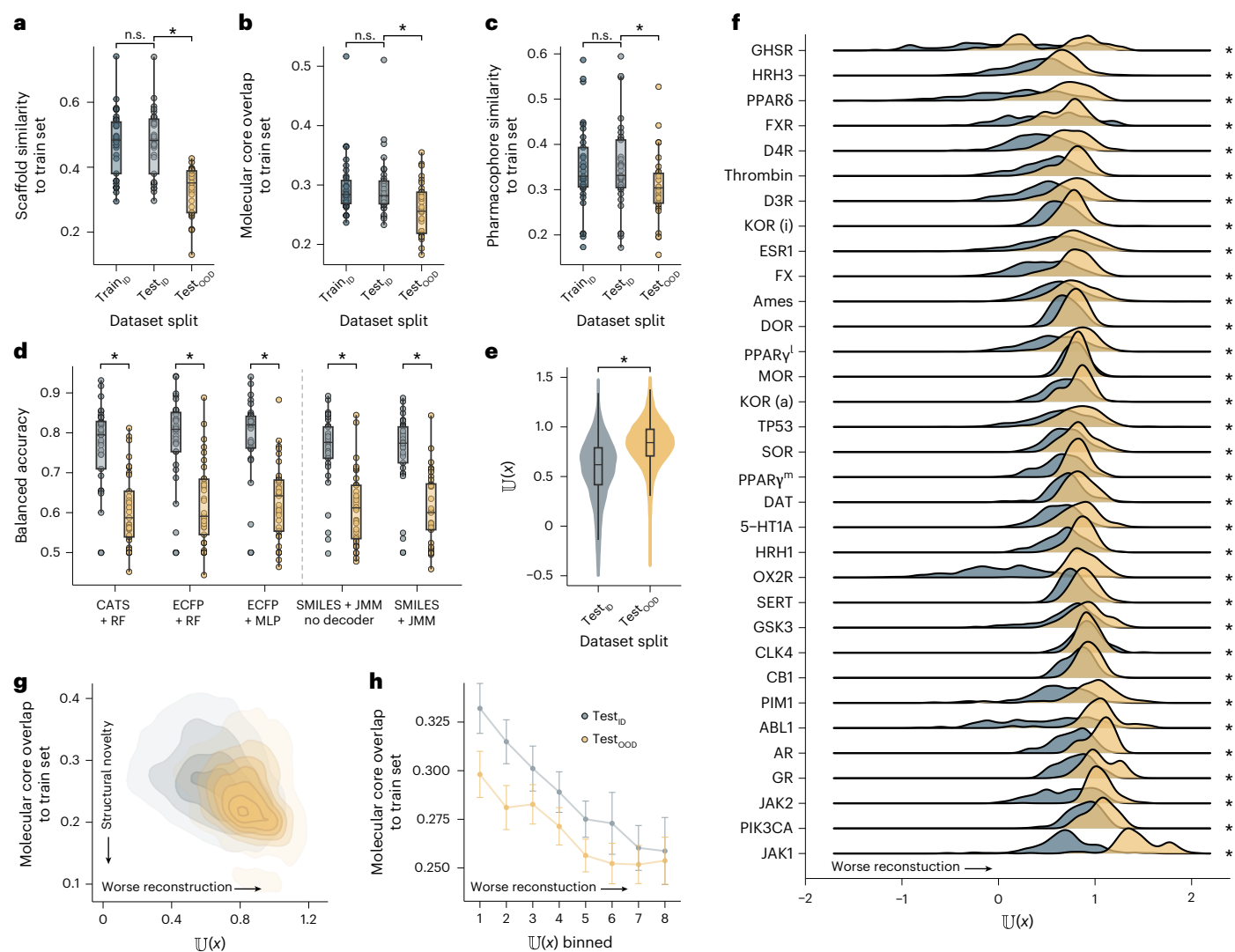


Fig. 2 | Detecting induced molecular distribution shifts with the unfamiliarity score. **a**, Mean scaffold similarity of data splits in the labelled datasets to their train set ($n = 33$ datasets). Similarity is calculated as the Tanimoto coefficient between ECFPs of Bemis–Murcko scaffolds. Every point in the box plot represents a dataset. **b**, Mean MCS fraction (MCSF) between data splits in the labelled sets to their train set ($n = 33$ datasets). **c**, Mean pharmacophore similarity (CATS¹²) between data splits in the labelled sets to their train set ($n = 33$ datasets). **d**, Predictive performance on the bioactivity finetuning sets ($n = 33$ datasets). From left to right, RF models using CATS descriptors, RF using ECFPs, MLP using ECFPs, MLP using a SMILES string encoder, and JMM using the same SMILES string encoder. **e**, Distribution of the JMM's unfamiliarity score for all molecules in test_{ID} and test_{OOD} across all labelled datasets ($n = 14,081$ molecules per split). Box plots in **a–e** show the median (centre line), 25th and 75th percentiles

(box bounds) and 1.5× interquartile range (IQR) (whiskers). **f**, Distributions of the JMM's unfamiliarity score in test_{ID} and test_{OOD} per datasets (dataset acronyms are specified in Supplementary Table 1). Statistically significant differences ($P < 0.05$) in **a–d** are denoted with an asterisk, determined by paired, two-sided, Wilcoxon signed-rank tests (**a–d**) and two-sided KS tests (**e** and **f**). Non-significant differences are denoted as n.s. **g**, Relationship between the JMM's unfamiliarity score and the mean MCSF similarity of all molecules in the labelled datasets ($n = 14,081$ molecules per split) to their respective train set. **h**, Relationship between binned unfamiliarity and MCSF similarities to the respective train set. Unfamiliarity values were binned per dataset. Points represent the mean over all datasets ($n = 33$), and error bars represent the standard error. All model-derived scores represent means over 10-fold Monte Carlo cross-validation (10% validation samples).

are low, are generally predicted accurately (Supplementary Fig. 4). Interestingly, unfamiliarity and prediction uncertainty seem to be unrelated as metrics themselves (Supplementary Fig. 5, Spearman correlation, $r = 0.10 \pm 0.05$). This aspect indicates that both metrics capture complementary information about prediction reliability and motivates the introduction of unfamiliarity alongside uncertainty for molecular machine learning.

'Model-driven' reliability metrics outperformed all 'data-driven' methods that calculate a molecule's distance to the training set using predefined descriptors and metrics. This indicates that models can extract relevant information from their training data that cannot be captured through predefined measures of molecular similarity alone.

Table 1 | Correlation of reliability metrics to train set similarity

| Reliability metric | Scaffold similarity | Molecular core overlap | Pharmacophore similarity |
|--------------------|---------------------|------------------------|--------------------------|
| Unfamiliarity | -0.46 ± 0.03 | -0.32 ± 0.04 | -0.24 ± 0.04 |
| Embedding distance | -0.36 ± 0.03 | -0.56 ± 0.02 | -0.10 ± 0.05 |
| Uncertainty | -0.15 ± 0.04 | 0.02 ± 0.04 | -0.15 ± 0.04 |

Spearman correlation between three model-based reliability metrics and several test-to-training set similarity metrics. Embedding distance is determined as the Mahalanobis distance of embeddings (\mathbf{x} vectors) to the training set. Mean and s.e.m. for all datasets are reported. Highest correlations are reported in bold.

Table 2 | Correlation of reliability metrics to model performance

| Binning metric | Rank-based correlation to bin order | | |
|--------------------------|-------------------------------------|------------------|------------------|
| | Balanced accuracy | Hit rate | Precision |
| Scaffold sim | 0.42±0.06 | 0.51±0.04 | 0.50±0.06 |
| Molecular core overlap | 0.28±0.07 | 0.22±0.09 | 0.25±0.07 |
| Pharmacophore similarity | 0.19±0.07 | 0.37±0.09 | 0.43±0.08 |
| Embedding distance | 0.36±0.06 | 0.24±0.09 | 0.29±0.08 |
| Uncertainty | 0.51±0.08 | 0.62±0.06 | 0.72±0.04 |
| Unfamiliarity | 0.58±0.04 | 0.52±0.07 | 0.52±0.05 |

Correlation (Kendall's τ) between several bin-wise performance metrics and the bin order. Molecules are binned into eight bins per dataset by: mean pharmacophore similarity to the training set (cosine distance computed on CATS descriptors), mean scaffold (Tanimoto on ECFPs) similarity to the training set, mean molecular core overlap (MCS fraction) to the training set, Mahalanobis distance of embeddings (\mathbf{z} vectors) to the training set, prediction uncertainty and unfamiliarity. Mean and standard error of the mean for all datasets are reported. A correlation of 1.0 indicates perfect model calibration. For every metric, bins are ordered to reflect low to high confidence. Highest correlations are reported in bold.

Interestingly, the unfamiliarity score proved to be a considerably better reliability metric than the embedding distance to the training embeddings, even though both capture elements of structural similarity learned by the JMM (Table 1). In other words, the ability of a model to reconstruct a molecule from an internal representation provides more insight into prediction reliability than the 'internal' molecular embeddings directly used by the classifier. This suggests that reconstructing a molecule from an embedding is not only highly informative for prediction reliability, but also for the quality of the embedding itself. These findings corroborate that simple embedding distance metrics fail to capture the 'chemical nuances' that affect task-specific outcomes and that embedding quality is better assessed through downstream tasks than by proximity to training embeddings, as previously suggested for computer vision⁵⁹.

Furthermore, our results highlight the distinctions between embedding distance, unfamiliarity and prediction uncertainty. While both unfamiliarity and prediction uncertainty correlate with performance independently, the latter does not strongly relate to the structural properties of molecules (Table 1). By contrast, unfamiliarity captures both molecular similarity and predictive performance simultaneously. In other words, embedding distance primarily reflects $p(x)$, uncertainty relates to $p(y|x)$, and unfamiliarity integrates both, providing insight into $p(y, x|x)$. This suggests unfamiliarity as a holistic metric of prediction reliability, effectively linking structural information to prediction confidence.

Virtual hit screening

To further explore the use of unfamiliarity to navigate chemical space, we extended our analysis to large-scale screening libraries to mimic a realistic virtual screening scenario. Although these libraries are unlabelled (and therefore do not allow performance evaluation), they contain a larger and more diverse set of molecules, potentially revealing additional differences compared with the smaller test sets analysed previously.

Using a combined set of 1.4 M molecules from three reliable commercial screening libraries (Asinex⁶⁰, Specs⁶¹ and Enamine⁶²; Supplementary Table 4), we performed inference with all 33 trained models. The screening molecules showed a lower structural overlap with the training sets than the test_{OOD} molecules (Fig. 3a and Supplementary Fig. 6a). Based on previous results (Fig. 2d), this indicates that we can also expect a corresponding performance drop on the screening molecules. Still, estimated prediction uncertainty did not highlight meaningful differences between screening molecules and test_{ID}

molecules (Fig. 3b; KS statistic $D = 0.181$, indicating a limited effect size). Prediction uncertainty alone would suggest that the screening libraries fall within a model's operating limits. Unfamiliarity scores, meanwhile, reveal strong distribution shifts, both overall (Fig. 3c; KS statistic $D = 0.999$), and especially on a dataset basis (Supplementary Fig. 6).

The estimated prediction uncertainty on this large-scale library does not directly correlate with a molecule's structural similarity to the training data (Supplementary Fig. 7a, Spearman correlation, $r = -0.04 \pm 0.02$). Unfamiliarity, meanwhile, shows a moderate relationship to structural similarity (Supplementary Fig. 7b, Spearman correlation, $r = 0.21 \pm 0.02$). Finally, prediction uncertainty and unfamiliarity remain independent (Fig. 3d, Spearman correlation, $r = -0.03 \pm 0.03$). Our findings confirm that the observed trends in small datasets remain robust across large screening libraries. Moreover, they highlight the ability of unfamiliarity to detect OOD shifts that routinely used metrics might fail to capture.

Uncertainty and unfamiliarity show stark differences across the molecular library, both from a pan-pharmacological angle (Fig. 3d) and for the well-studied serine/threonine-protein kinase PIM1 as a highlighted case (Fig. 3e). Globally, molecules with high $\mathbb{U}(x)$ scores are 'structurally atypical' (for example, molecules i and ii, Fig. 3d). Molecules with low $\mathbb{U}(x)$ scores, meanwhile, display key characteristics of bioactive molecules, for example, steroid-like structures⁶³ and bioactive cores⁶⁴ (molecules iii and iv, Fig. 3d). Similar trends are observed across individual protein targets, as exemplified by PIM1 (Fig. 3e). Here, molecules with low $\mathbb{U}(x)$ contain well-known pharmacophores (for example, pyrimidinone cores⁶⁵).

Finally, we found no relationship between unfamiliarity and the quantitative estimate of drug-likeness⁶⁶ (Spearman correlation: $r = -0.04 \pm 0.03$) or synthetic accessibility⁶⁷ (Spearman correlation: $r = 0.01 \pm 0.02$), demonstrating that low unfamiliarity scores do not simply capture drug-likeness. Importantly, structural insights are not captured by uncertainty estimation. This might indicate that uncertainty estimation is unreliable on strongly OOD data, as has been suggested previously^{14,24,25}.

Prospective virtual screening

To validate the proposed approach prospectively under real-world conditions, we screened a commercial compound library comprising approximately 180,000 drug-like molecules (from Specs⁶¹; Methods) to identify inhibitors of two pharmacologically relevant kinase targets: PIM1 and cyclin-dependent kinase 1 (CDK1). Notably, CDK1 data were not used elsewhere in this study, serving as a fully independent test case.

For each target, we trained a JMM on all available data (1,443 training molecules for PIM1 and 312 for CDK1) and predicted bioactivity across the screening library. Molecules were then ranked by their distance to the so-called utopia point⁶⁸—the geometric optimum balancing multiple objectives (Methods). Using the predicted bioactivity $\mathbb{E}(y|x)$, prediction uncertainty $\mathbb{H}(y|x)$ and unfamiliarity $\mathbb{U}(x)$ as three complementary objectives, we selected the ten best molecules with three alternative trade-offs to illustrate how uncertainty and unfamiliarity behave under distribution shifts (Fig. 4a,d):

- (A) High $\mathbb{E}(y, |, x)$, High $\mathbb{H}(y|x)$, and low $\mathbb{U}(x)$.
- (B) High $\mathbb{E}(y, |, x)$, Low $\mathbb{H}(y|x)$, and low $\mathbb{U}(x)$.
- (C) High $\mathbb{E}(y, |, x)$, Low $\mathbb{H}(y|x)$, and high $\mathbb{U}(x)$.

To limit structural overlap with the training data and among selected molecules, compounds with a Tanimoto coefficient (on ECFPs) ≥ 0.70 to either the training set or other selected compounds were excluded, thereby further challenging the models beyond their training distribution.

All 60 prioritized molecules (Supplementary Figs. 9 and 10) were structurally distant from their respective training sets, with

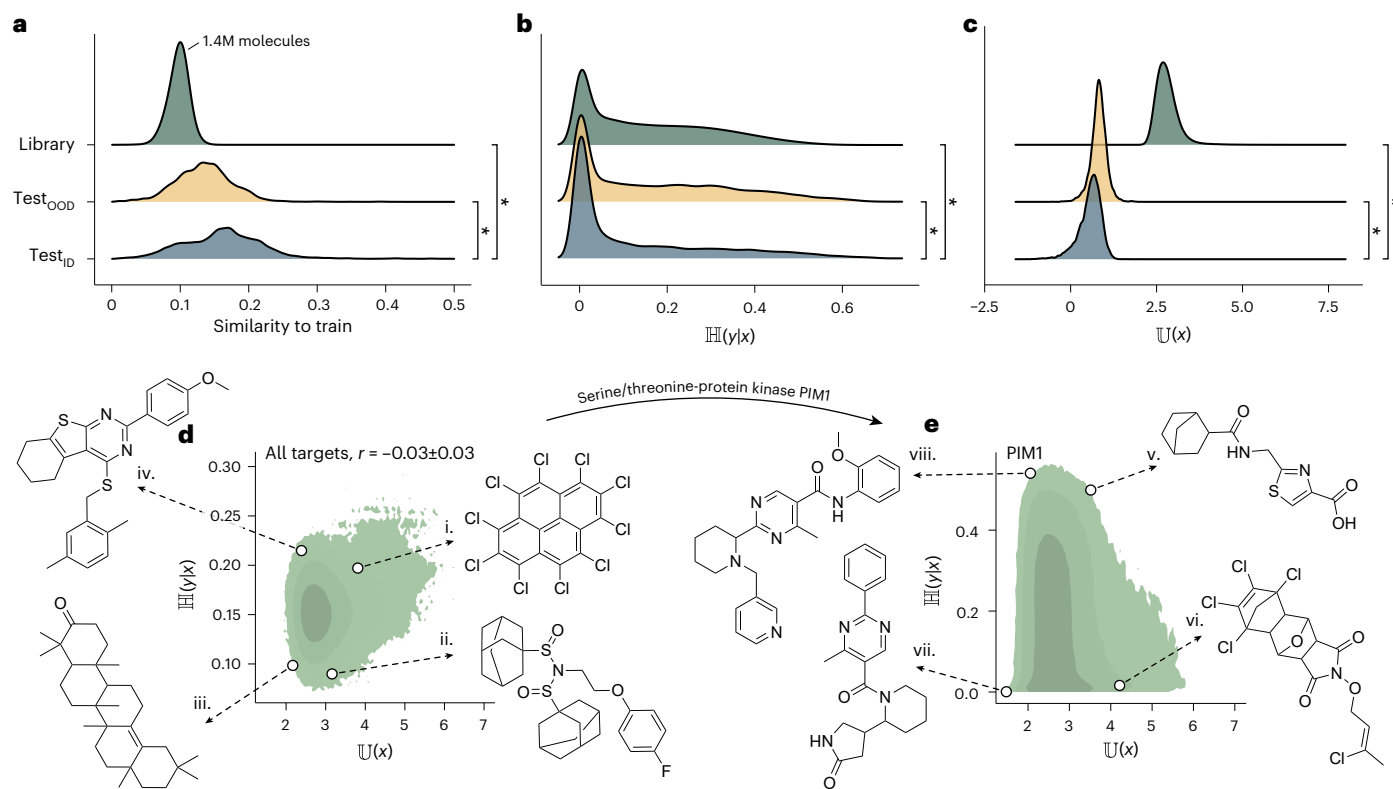


Fig. 3 | Virtual screening on 1.4M commercially available molecules. All model-derived scores represent means over 10-fold Monte Carlo cross-validation (10% validation samples). **a**, Distributions of the mean Tanimoto similarity on ECFPs to each respective training set of molecules from test_{ID} ($n = 14,081$), test_{OOD} ($n = 14,081$) and the combined screening libraries ($n = 46,048,926$). Results of all 33 drug targets are combined. **b**, Distributions of estimated prediction uncertainty $\mathbb{H}(y|x)$ for all molecules in test_{ID} , test_{OOD} and the combined screening libraries. **c**, Distributions of unfamiliarity scores $\mathbb{U}(x)$ for all molecules in test_{ID} , test_{OOD} and the combined screening libraries. Statistically significant differences ($P < 0.001$) are denoted with an asterisk, determined by two-sided KS tests (**a-c**). KS test statistics (D) are as follows. **a**: Library versus test_{ID} , $D = 0.499$; test_{OOD}

versus test_{ID} , $D = 0.279$; **b**: library versus test_{ID} , $D = 0.181$; test_{OOD} versus test_{ID} , $D = 0.155$; **c**: library versus test_{ID} , $D = 0.999$; test_{OOD} versus test_{ID} , $D = 0.368$. **d**, Relationship between uncertainty and unfamiliarity for all molecules in the screening libraries ($n = 1,395,422$), averaged over all 33 drug targets. The mean Spearman correlation is reported \pm s.e.m. Four molecules predicted as generally bioactive across all drug targets are annotated, each close to a utopia point (for example, lowest uncertainty and lowest unfamiliarity for molecule iii; Methods). **e**, Relationship between uncertainty and unfamiliarity for all molecules in the screening libraries, specifically for serine/threonine-protein kinase PIM1 ($n = 1,395,422$). Four molecules predicted as bioactive for PIM1 are annotated, each close to a utopia point.

maximal Tanimoto coefficients (on ECFPs) of 0.28 ± 0.05 for PIM1 and 0.28 ± 0.06 for CDK1. Molecules selected via the high unfamiliarity strategy were particularly atypical for kinase inhibitors (Supplementary Figs. 9 and 10), aligning with previous findings (Fig. 3d,e). After experimentally testing all compounds at a single concentration of $10 \mu\text{M}$ (Fig. 4b and Supplementary Fig. 8), we identified four initial hits ($>50\%$ protein inhibition) and six weak hits ($>25\%$ inhibition) for PIM1. For CDK1, we found one initial hit and five weak hits.

The six most active compounds per target were further characterized to determine dose–response curves and the corresponding half maximal inhibitory concentration (IC_{50}). For PIM1, all six compounds showed dose-dependent inhibition (Supplementary Fig. 11), although none achieved full inhibition within the tested concentration range ($1 \text{ nM} - 10 \mu\text{M}$). Compounds **4**, **10** and **25** exhibited clear sigmoidal partial inhibition curves, with low/sub micromolar potency (IC_{50} $1.5 \pm 0.4 \mu\text{M}$, IC_{50} $0.92 \pm 0.97 \mu\text{M}$ and IC_{50} $0.87 \pm 0.5 \mu\text{M}$, respectively). Compounds **9** and **18** demonstrated weaker inhibition with incomplete curve resolution; for these, we report upper-bound potency estimates (IC_{50} $5.9 \mu\text{M}$ and IC_{50} $7.5 \mu\text{M}$, respectively).

For CDK1, inhibition was generally less pronounced (Fig. 4d and Supplementary Fig. 12), possibly due to the smaller training set (312 molecules). Only compound **40** yielded a complete dose–response curve, with partial inhibition (IC_{50} $2.9 \pm 0.75 \mu\text{M}$). Compound **33** showed modest activity with an upper-bound IC_{50} $\sim 3.4 \mu\text{M}$. The remaining

compounds (**31**, **48**, **53** and **59**) showed insufficient inhibition to estimate an IC_{50} .

Overall, our screening experiment achieved hit rates of approximately 17% for PIM1 (5 of 30 compounds with clear dose–response inhibition) and 7% for CDK1 (2 of 30), with several additional weak actives. These hit rates exceed those typically reported for traditional kinase-focused screening campaigns, which often range from 0.1% to 5% (refs. 69–72). Notably, all identified hits had a maximal substructure similarity to the training molecules below 38% (measured as the Tanimoto coefficient on ECFPs), and emerged from purely prospective, machine learning-guided selection.

While the limited number of compounds tested per strategy and the absence of a full 2×2 factorial design prevents us from drawing statistically robust conclusions and cleanly separating main effects, five of the seven compounds with low micromolar potency originated from selection method A (low unfamiliarity, high uncertainty). By contrast, selecting molecules with low prediction uncertainty (methods B and C) did not yield a clear advantage, consistent with earlier results indicating that uncertainty is not a reliable performance signal under distribution shifts (Fig. 3a–c). Although preliminary, these prospective results provide practical evidence that the proposed approach can identify novel bioactive matter and support unfamiliarity as a useful handle for navigating chemical space under distribution shifts.

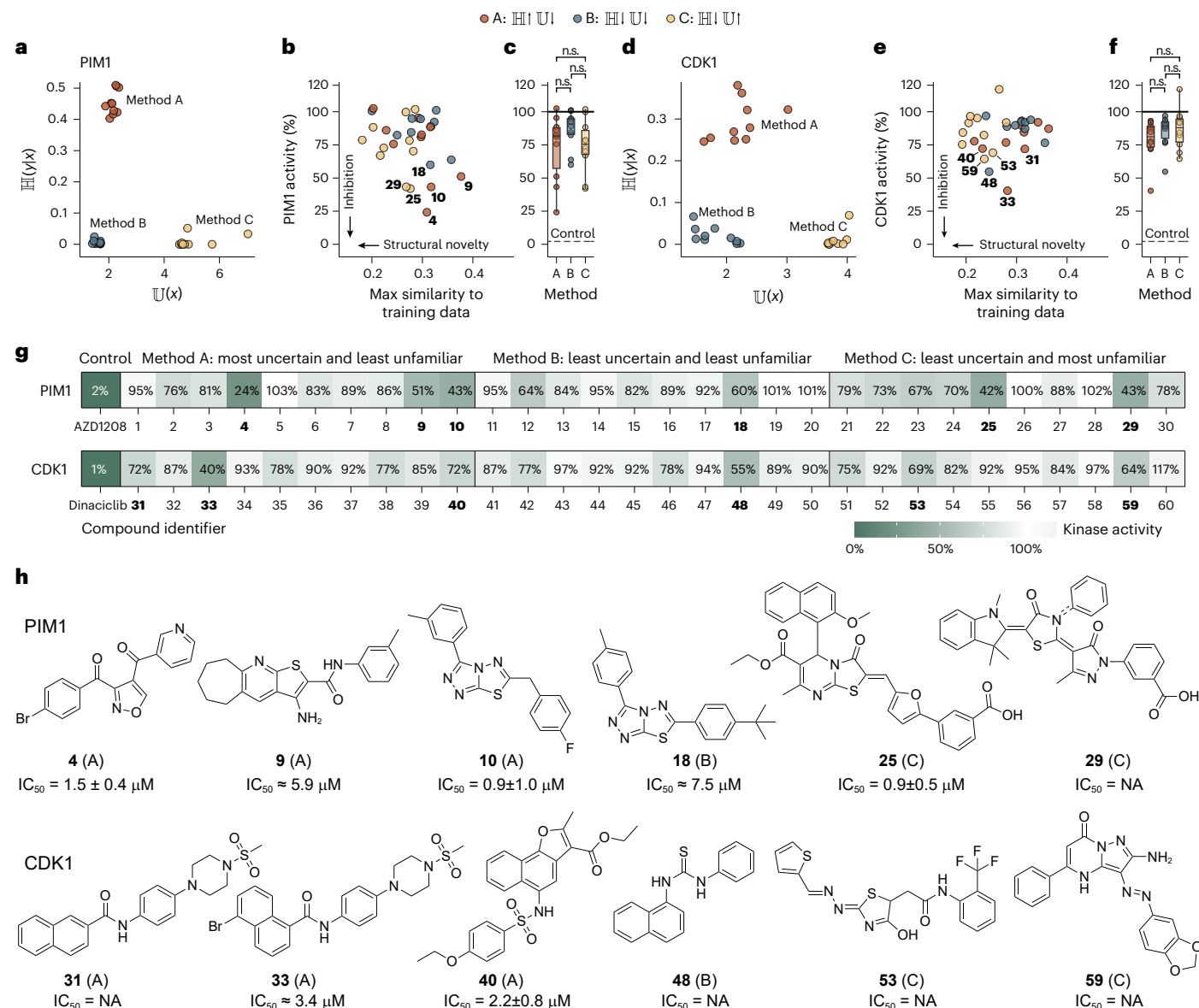


Fig. 4 | Experimental screening for PIM1 and CDK1. Compound selection was based on unfamiliarity scores and uncertainty estimates averaged over 10-fold Monte Carlo cross-validation (10% validation samples). All kinase activity measurements represent the mean of three technical replicates, except for positive control compounds (AZD1208 and dinaciclib). **a**, The ten most promising PIM1 inhibitors were selected from a library of ~180,000 compounds using three combinations of uncertainty and unfamiliarity (A: most uncertain and least unfamiliar; B: least uncertain and least unfamiliar; C: least uncertain and most unfamiliar). **b**, Measured PIM1 activity across selected molecules at 10 μM and their maximum Tanimoto similarity (on ECFPs) to PIM1 training molecules. Lower PIM1 activity means stronger inhibition. **c**, Box plot of measured PIM1 activity ($n = 10$ molecules per method). The solid line represents PIM1 activity without any screening compound, while the dashed line represents PIM1 activity with a potent control inhibitor (AZD1208). Statistically significant differences ($\alpha = 0.05$) are denoted with an asterisk, and were determined by paired, two-sided, Wilcoxon signed-rank tests. Box plots show the median (centre line),

25th and 75th percentiles (box bounds) and $1.5 \times IQR$ (whiskers). **d**, The ten most promising CDK1 inhibitors for each selection method. **e**, Measured CDK1 activity across selected molecules at 10 μM and their maximum Tanimoto similarity (on ECFPs) to CDK1 training molecules. **f**, Box plot of measured CDK1 activity ($n = 10$ molecules per method). The solid line represents CDK1 activity without any screening compound, while the dashed line represents CDK1 activity with a potent control inhibitor (dinaciclib). Box plots show the median (centre line), 25th and 75th percentiles (box bounds) and $1.5 \times IQR$ (whiskers). Statistically significant differences ($\alpha = 0.05$) are denoted with an asterisk, and were determined by paired, two-sided, Wilcoxon signed-rank tests. **g**, Measured protein activity across all 60 screened compounds at 10 μM (method A: 1–10 and 31–40; method B: 11–20 and 41–50; method C: 21–30 and 51–60). Selected compounds (identifier highlighted in boldface) displayed in panel **h** were further characterized for their dose–response curve. **h**, Structures and determined IC_{50} of the six most promising compounds for PIM1 and CDK1. IC_{50} values that could not be determined are denoted as NA.

Discussion

This study introduced unfamiliarity—a metric that captures a molecule's distance from the data distribution learned by a deep learning model. Unfamiliarity is computed via a joint modelling approach, trained to simultaneously perform molecular reconstruction and property prediction. By capturing the model error in reconstructing previously

unseen molecules, unfamiliarity quantifies molecular distribution shifts. Our results demonstrate that unfamiliarity is a reliable and powerful indicator of a model's prediction reliability when applied to new molecules. As a classification reliability metric, unfamiliarity is as informative as prediction uncertainty estimated via approximate Bayesian modelling, yet the two remain independent. Unlike

prediction uncertainty, however, unfamiliarity also captures a molecule's structural distance from the learned data distribution. Notably, in a large-scale virtual screening campaign, unfamiliarity provided far more meaningful molecular insights than uncertainty estimation when faced with strong distribution shifts. The complementarity of unfamiliarity and prediction uncertainty as reliability metrics highlights unfamiliarity as a valuable tool for molecular machine learning.

The prospective validation further underscores the usefulness of the introduced unfamiliarity metric to complement uncertainty-based molecule prioritization. Despite using small training sets and screening only a handful of compounds per target, we identified multiple molecules with low micromolar activity. Prediction uncertainty, meanwhile, did not seem indicative of performance when operating outside the training support^{14,24,25}. This suggests that unfamiliarity-aware selection can enable informed and precise exploration of the chemical space, unlocking new opportunities for discovering novel molecular candidates.

Our findings advocate for the adoption of unfamiliarity over traditional, similarity-based methods for applicability domain definition. Moreover, because unfamiliarity bears promise to reveal distribution shifts that would be undetected through molecular similarity or underestimated by uncertainty-based approaches, we recommend its adoption when screening large-scale molecular libraries. Crucially, because the unfamiliarity landscape reveals gaps in a model's learned distribution, its applications could assist in guiding reinforcement learning applications⁷³ or extend beyond 'one-shot' virtual screening to iterative approaches such as active learning^{74,75} to guide molecule acquisition. Ultimately, our study highlights the advantages of joint modelling not only in de novo design applications²⁷ but also in capturing structure–activity relationships. The information provided by joint modelling and unfamiliarity is expected to drive the development of more reliable and generalizable models—accelerating the exploration of novel regions in the chemical space with greater confidence and precision.

Methods

Data preprocessing and analysis

Molecular representation and description. Each molecular structure in this study was represented as a SMILES⁴² string. CATS⁵⁶ descriptors and 2048-bit ECFPs⁵³ (using a radius of 2) were computed for each molecule.

Data collection and curation. Thirty-three labelled datasets of molecular structures with their corresponding experimental target property were used:

- LIT-PCBA⁴⁹ (3 targets). The ESRI (antagonism), TP53 and PPAR γ bioactivity datasets were downloaded from LIT-PCBA⁴⁹ (accessed in August 2023 at <https://drugdesign.unistra.fr/LIT-PCBA>).
- The Ames mutagenicity dataset⁵⁰ was downloaded from <http://pubs.acs.org>.
- MoleculeACE⁴⁸ (29 targets). Bioactivity datasets were downloaded from <https://github.com/molML/MoleculeACE>.

Moreover, small molecules were collected from ChEMBLv33⁴⁵ for model pretraining. Because molecular structures from ChEMBL were used for pretraining, molecules with a Bemis–Murcko scaffold⁷⁶ similar to any such scaffold in the labelled datasets (Tanimoto similarity coefficient⁵⁴ on EFCPs larger than 0.7) were removed. This included molecules without a Bemis–Murcko scaffold (that is, containing no rings).

For prospective virtual screening, the most recent CDK1 (ChEMBL308) data were fetched from ChEMBL v35⁴⁵ (accessed in April 2025) as an additional and independent dataset. Raw data were processed in accordance with previous work⁴⁸.

All bioactivity endpoints of the MoleculeACE⁴⁸ and CDK1 datasets were converted from continuous regression labels into binary

classification labels. Molecules with an EC₅₀ (half maximal effective concentration) or K_i (inhibitory constant) of 100 nM or lower were labelled as bioactive, whereas less potent molecules were labelled as inactive. For the LIT-PCBA⁴⁹ and Ames mutagenicity dataset⁵⁰, their original binary classification labels were used.

For all datasets, SMILES strings were preprocessed using RDKit v. 2024.3.3⁷⁷. For each SMILES string, stereochemistry tokens as well as salts and solvents (Supplementary Table 2) were removed. Each molecule was sanitized, neutralized using predefined neutralization reactions (Supplementary Table 3), and its SMILES string was canonicalized. Disconnected structures and molecules that contained formal charges, contained complex ring systems (SMILES strings with a ring index of 9 or higher), non-standard isotopes or any atoms other than Cl, Br, H, C, N, O, F, S and I were removed. Molecules were removed if they contained more than 100 tokens in their canonical SMILES string or if they could not be featurized into CATS descriptors and/or ECFP fingerprints. Sizes of datasets before and after data curation are presented in Supplementary Table 1.

Virtual screening library (retrospective). Molecules were gathered from three commercial screening libraries:

- Asinex⁶⁰ screening libraries: downloaded from [https://www.asinex.com/screening-libraries-\(all-libraries\)](https://www.asinex.com/screening-libraries-(all-libraries)) (accessed in March 2025).
- Specs⁶¹: downloaded from <https://www.specs.net> (accessed in March 2025).
- Enamine hit locator⁶²: downloaded from <https://enamine.net/compound-libraries/diversity-libraries/hit-locator-library-460> (accessed in March 2025).

All molecules were processed in the same way as the training datasets. Finally, all unique molecules were aggregated.

Virtual screening library (prospective). The most recent Specs library was downloaded from https://www.specs.net/index.php?view=data_bases&page=download (accessed in April 2025). All molecules were processed in the same way as the training datasets. To ensure that molecules were compatible with our experimental assay (likely soluble in 1% dimethylsulfoxide (DMSO)), physicochemical rules were enforced based on relaxed rule-of-five and Veber criteria:

- (1) Molecular weight between 200 g mol⁻¹ and 600 g mol⁻¹.
- (2) log P lower than 6.
- (3) Total polar surface area between 20 Å² and 140 Å².
- (4) Number of hydrogen bond donors lower than six.
- (5) Number of rotatable bonds lower than ten.
- (6) A maximum of two rule-of-five violations.

In addition, molecules with a terminal enone, isocyanate, quinone, aromatic nitro groups, azide or epoxide groups were removed by using SMILES Arbitrary Target Specification (SMARTS) patterns to prevent assay interference. Furthermore, to prevent the selection of trivially simple molecules and enrich the general-purpose Specs library for kinase-relevant chemical space, several general criteria were enforced:

- (1) A molecule must have an ATP mimetic core, that is, at least one heteroatom in a ring, or a fused carbocyclic system.
- (2) A molecule must have a polar anchor to ensure solubility or solvent interaction.
- (3) A molecule must have enough hydrophobic mass or planarity to fill the kinase ATP pocket (for example, gatekeeper region).
- (4) A molecule must have a directional H-bond donor/acceptor to ensure potential for hinge interaction.
- (5) A molecule must have two or more rings.

These rules are intentionally permissive for new chemotypes, and 98.6% of all kinase inhibitors that ever went into clinical trials⁷⁸ pass

these filters. Finally, molecules with a Tanimoto similarity on ECFPs >0.7 to any molecule in the respective target's data (PIMI, and CDK1) were removed. This left 185,298 and 185,336 screening molecules for PIMI and CDK1, respectively.

Molecular cyclic skeletons. Cyclic skeletons⁵¹ (core ring systems without exocyclic substituents) were extracted from molecules to serve as the most fundamental molecular scaffold representation. From each molecule, Bemis–Murcko scaffolds⁷⁶ were obtained, removing peripheral substituents. Remaining double-bonded exocyclic substituents were then removed, and all atoms and bonds were made generic to obtain the final cyclic skeleton.

Data splitting. The curated molecular structures from ChEMBL were split into a training (80%, $n = 1,230,041$), a test set (10%, $n = 153,755$) and a validation (10%, $n = 153,755$) set, using random splitting. All labelled datasets were split into a training set (~50%), test set (~25%) and OOD set (~25%). To determine the OOD molecules, spectral clustering was performed on unique cyclic skeletons (see below). The molecules corresponding to the scaffolds in the clusters with the lowest mean cluster similarity to all other clusters that constituted approximately 25% of the total dataset size were taken as the OOD set. The remaining molecules (that is, the ~75% most similar molecules) were split randomly in a train and test set, with the test set being equal in size to the OOD set. An overview of all datasets is presented in Supplementary Table 1.

Spectral clustering

Spectral clustering was performed on a molecular similarity matrix A using Sci-kit learn⁷⁹. A is an $n \times n$ matrix where each element A_{ij} is the Tanimoto coefficient on ECFPs T_{ij} between two molecular structures. By using a molecular similarity matrix directly, we bypass the complex, high-dimensional and non-Euclidean nature of molecular structures. From this affinity matrix, the symmetrically normalized Laplacian was constructed as follows:

$$L_{\text{sym}} = I - D^{-1/2}AD^{-1/2}, \quad (3)$$

where I is the identity matrix and D is the degree matrix. Subsequently, eigenvalue decomposition was performed:

$$L_{\text{sym}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T. \quad (4)$$

To determine the number of clusters k for spectral clustering, the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ were sorted in ascending order and the elbow (the point of maximal curvature) of the resulting sequence was estimated using the kneed algorithm⁸⁰. Finally, the spectral embeddings of the data were clustered by taking the top k (smallest) eigenvectors \mathbf{U}_k , normalizing \mathbf{U}_k to unit length and performing k -means clustering on all rows (u_k).

Chemical space visualization. To visualize each labelled dataset, molecules were first encoded as ECFPs. The resulting binary ECFPs were reduced to 100 components using truncated singular value decomposition and embedded into a two-dimensional space using t -distributed stochastic neighbour embedding with a perplexity value of 30 and default settings.

MCS fraction. To compute molecular core similarity, we computed the MCS fraction⁵⁵ between a molecule M_a and a reference molecule M_b as

$$\text{MCS fraction} = \frac{|\text{MCS}(M_a, M_b)|}{|M_a|}, \quad (5)$$

where $|M_a|$ is the number of atoms in M_a and MCS is the maximal common substructure between the two molecules M_a and M_b , as determined by the FMCS algorithm⁷ in RDKit. A high MCS fraction indicates that a

molecule shares a significant portion of their overall core structure with a reference molecule. This implementation is asymmetric.

Molecular complexity. To quantify molecular complexity we compute the well-established Bertz complexity⁸¹ and Böttcher complexity⁸². In addition, to align complexity measures with the task of reconstructing SMILES strings, we also estimated the complexity of the molecular graph directly⁸³ as

$$C_{\text{graph}} = V \log_2 V - \sum_{i \in \mathcal{G}} V_i \log_2 p_i, \quad (6)$$

where G represents the molecular graph, V is the total number of elements in the graph and V_i is the number of the elements in the i th set of elements. In a similar fashion, we also estimated the complexity of a SMILES string, via their entropy computed as

$$C_{\text{SMILES}} = - \sum_{i \in S} p_i \log_2 p_i, \quad (7)$$

where S represents the set of unique tokens in a SMILES string, and p_i is the probability of the i th token occurring in S . Tokens representing the start, end of sequence, and padding were not considered. Moreover, for each molecule, we counted the number of SMILES tokens (excluding padding), the number of SMILES string branches (that is, '(' tokens), and the presence of 50 unique molecular patterns⁷⁴.

Virtual screening

Top- k molecules were selected in a multi-objective manner by selecting the k molecules closest to the utopia point⁶⁸. The distance to this ideal point can be calculated as

$$d_{\text{utopia}} = \sqrt{\sum_{i=1}^n (\text{norm}_i)^2}, \quad (8)$$

where norm_i is the normalized objective that is either maximized (for example, predicted bioactivity) or minimized (for example, prediction uncertainty):

$$\text{norm}_i^{\text{[max]}} = \frac{x_{\text{max}} - x_i}{x_{\text{max}} - x_{\text{min}}} \quad \text{or} \quad \text{norm}_i^{\text{[min]}} = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}. \quad (9)$$

Machine learning

Encoder. Canonical SMILES strings were encoded by a one-dimensional convolutional neural network. SMILES string character tokens were embedded using a randomly initialized trainable embedding layer of size 128. Several one-dimensional convolutional layers were used with a stride of 1 and no padding. Each layer was followed by a ReLU activation, standard max pooling with a kernel size equal to that of the convolutional layers, and dropout. Both convolutional and pooling layers used a stride of 1 and no padding. The final output was flattened and compressed to a latent vector (\mathbf{z}) of size 128 using a fully connected layer. The following hyperparameters were optimized (see 'Hyperparameter optimization' section): the number of convolutional layers [2, 3], filter size [256, 512], kernel size [6, 8], weight decay on convolutional neural network weights [$0, 1 \times 10^{-4}$] and dropout [0, 0.1].

Decoder. Encoded latent molecular representations in \mathbf{z} were reconstructed back to SMILES strings using a conditioned LSTM model. A randomly initialized trainable embedding layer of 128 neurons was used to embed SMILES string character tokens. The following hyperparameters were optimized (see 'Hyperparameter optimization' section): the number of LSTM layers (n_{layers}) [2, 3] and the LSTM hidden size ($\text{size}_{\text{layers}}$) [256, 512]. Models were trained autoregressively without teacher forcing using next token prediction based on the tokens

predicted in the previous steps rather than the ground truth tokens. A reconstruction loss normalized for sequence length was used (equation (1)). To condition the model, the LSTM hidden state h_0 was initialized with \mathbf{z} for every molecule. To correctly match all dimensions of h_0 ($n_{\text{layers}}, \text{size}_{\text{layers}}$), \mathbf{z} was first transformed to $n_{\text{layers}} \times \text{size}_{\text{layers}}$ with a fully connected layer, after which it was reshaped into n_{layers} chunks of $\text{size}_{\text{layers}}$.

Approximate Bayesian classifier. Labels were predicted from either ECFPs or SMILES strings encoded into latent molecular representations \mathbf{z} . A MLP was used with several fully connected layers and an output layer consisting of two neurons. The number of MLP layers [2, 3] and the MLP hidden size [1,024, 2,048] were optimized (see ‘Hyperparameter optimization’ section). To estimate prediction uncertainty, anchored ensembling⁴⁷ was implemented on the MLP as in our previous work⁷⁴. We used an ensemble of $M = 10$ models. For each model, $m \in [1 \dots M]$, we regularized its parameters θ_m with a set of ‘anchored’ parameters $\theta_{\text{anchor},m}$ that prevent different models in the ensemble to converge to the same parameter space. Each model is initiated with distinct θ_{anchor} , which is controlled by different random seeds. The classification loss in our implementation is defined as

$$\mathcal{L}_{\text{MLP}} = -\frac{1}{M} \sum_{m=1}^M \left(\underbrace{\log p_m(y|x)}_{\text{prediction}} + \lambda \underbrace{\|\theta_m - \theta_{\text{anchor}}^m\|^2}_{\text{anchoring}} \right), \quad (10)$$

where λ is a regularization coefficient (set to 3×10^{-4}). To estimate the expected value \mathbb{E} of each molecule x , we take the mean prediction over the whole ensemble, as follows:

$$\mathbb{E}(y|x) = \frac{1}{M} \sum_{m=1}^M p_m(y, |, x). \quad (11)$$

Similarly, we estimate prediction uncertainty for each molecule x as the mean entropy \mathbb{H} over all models in the ensemble:

$$\mathbb{H}(y|x) = -\frac{1}{M} \sum_{m=1}^M p_m(y|x) \log p_m(y|x). \quad (12)$$

JMM. Canonical SMILES strings were encoded into compressed latent molecular representations \mathbf{z} using the encoder model described above. Subsequently, \mathbf{z} was used to perform both molecular property prediction with a classifier and molecular reconstruction using the \mathbf{z} -conditioned decoder model²⁷. The model was optimized in a joint fashion using the following weighted composite loss function:

$$\mathcal{L}_{\text{JMM}} = \mathcal{L}_{\text{reconstruction}} + \gamma \times \mathcal{L}_{\text{MLP}}, \quad (13)$$

where the scalar γ was set at 0.1 based on preliminary experiments. A regular autoencoder was used as preliminary experiments showed no performance benefits of the more complex variational autoencoder⁸⁴ architecture.

RF. An RF classifier was trained on either ECFPs or CATS⁵⁶ descriptors. The following hyperparameters were optimized (see ‘Hyperparameter optimization’): the number of trees [100, 250, 500, 1,000], the maximal tree depth [10, 20, 30, ∞] and the minimal samples per split [2, 5, 10].

Model training

Autoencoders. Encoder–decoder models were (pre)trained to reconstruct SMILES strings of general drug-like molecules from ChEMBL using the Adam optimizer. Mini batches of 256 random molecules were sampled from the training data for 1,000,000 steps using uniform sampling. Gradients were clipped with a max norm of 5. Early stopping with a patience of up to 20 evaluation checkpoints was implemented by

monitoring validation loss every 10,000 steps. The model checkpoint with the best validation loss was used.

Classifiers. Classifiers using ECFPs or SMILES strings as input were trained for molecular property prediction on each of the labelled dataset using a similar setup to the autoencoders. However, tenfold Monte Carlo cross-validation was used with 10% validation splits instead of one predefined data split. Mini batches of 64 were resampled during training based on the occurrence of their class with

$$P_c = 1 - \frac{n_c}{N}, \quad (14)$$

where P_c is the probability of sampling class c , n_c is the number of samples of class c , and N is the total number of samples. Models were trained for 5,000 steps with an early stopping patience of 10 evaluation checkpoints, performed every 10 steps. The model checkpoint with the best validation loss was used. For the RF control models, molecules were weighted inversely proportionally to their class frequency to mitigate class imbalance during training.

JMMs. Joint models, each consisting of a SMILES string encoder, a decoder and a classifier (Fig. 1b), were initialized with pretrained weights. The SMILES string encoders and classifiers used weights from models trained on the labelled datasets. For the decoder, decoder weights were used from an autoencoder pretrained on ChEMBL. Using a mini batch size of 64, the joint models were finetuned for 10,000 steps with an early stopping patience of 50 evaluation checkpoints, performed every 20 steps. A learning rate of 3×10^{-6} was used for the encoder and classifier, whereas a learning rate of 3×10^{-7} was used for the decoder. No weight decay was applied.

Hyperparameter optimization. Hyperparameters, as specified previously, were optimized for all autoencoders, classifiers (using ECFPs or SMILES strings), and RF control models using a simple grid search. Tenfold Monte Carlo cross-validation was used, repeatedly using 10% of the training data as a validation split. The hyperparameters of the model with the best mean validation loss was used.

Model evaluation

Predictions were evaluated according to the following performance metrics:

$$\text{Balanced accuracy} = \frac{1}{2} \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right), \quad (15)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (16)$$

where TP is the number of true (that is, correctly predicted) positives, TN is the number of true negatives, FN is the number of false negatives, and FP is the number of false positives. In addition, the hit rate (true positive rate) was determined for virtual screening experiments as

$$\text{Hit rate} = \frac{\text{TP}}{P}, \quad (17)$$

and enrichment factor as

$$\text{Enrichment factor} = \frac{\text{TP}_k}{P/N}. \quad (18)$$

Here, TP_k is the number of correctly identified positives in the subset of k prioritized molecules, P is the total number of positives in the full dataset, and N is the total number of molecules in the full dataset.

Biological characterization

Sixty screening compounds were purchased from Specs Compound Handling B.V. and dissolved at 10 mM in 100% DMSO.

To screen for bioactivity, a point screening was first performed at a concentration of 10 μ M (in 1% DMSO) in technical triplicates using the ADP-Glo Kinase Assay platform from Promega using the Chemi-Verse PIM1 Kinase Assay Kit and the Chemi-Verse CDK1/CyclinA2 Kinase Assay Kit from BPS Bioscience in Costar flat white 96-well plates. AZD1208 (CAS 1204144-28-4) and dinaciclib (CAS 779353-01-4), purchased from TargetMol Chemicals, were used as positive controls for PIM1 and CDK1, respectively. Bioactivity was measured as the area under the curve of an 18-step luminescence scan between 398 nm and 653 nm with an integration time of 1 s and a settle time of 100 ms, normalized for the signal in buffer-only wells.

For each target protein, the six compounds with the highest bioactivity at 10 μ M were followed up with an 8-point dose–response curve using the same assay. Screening compounds were measured in technical triplicates from 10 μ M to 0.0046 μ M, whereas reference compounds were measured in duplicate.

Hardware and training set-up

All computational experiments were performed on a Lenovo ThinkSystem SD650-N v2 server equipped with Intel Xeon Platinum 8360Y central processing units and NVIDIA A100 (40 GB) graphics processing units. Up to five models were trained in parallel on a single graphics processing unit.

Software and code

All code was implemented in Python (v. 3.12). Deep learning models were implemented using PyTorch (v. 2.3.0)⁸⁵. Traditional machine learning models and clustering was implemented using Sci-kit learn v.1.4.0 (ref. 79). All molecular data were handled using RDKit (v.2024.3.3)⁷⁷. For data visualization, R (v.4.3.0) and the R package ggplot2 (v.3.4.2) were used along with Adobe Illustrator.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All processed datasets and results are available via Zenodo at <https://doi.org/10.5281/zenodo.14865513> (ref. 86), except for data derived from commercial screening libraries due to licensing restrictions. Source data are provided with this paper. These data are also available via figshare at <https://doi.org/10.6084/m9.figshare.30665201.v1> (ref. 87). Source data are provided with this paper.

Code availability

Code is available via figshare at <https://doi.org/10.6084/m9.figshare.30665201.v1> (ref. 87). The Python code to replicate and extend our study, alongside the R scripts to visualize all results, is available via GitHub at <https://github.com/molML/JointMolecularModel> and via Zenodo at <https://doi.org/10.5281/zenodo.18846066> (ref. 88).

References

- Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. *Mol. Inform.* **35**, 3–14 (2016).
- Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin. Drug Discov.* **16**, 949–959 (2021).
- Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
- Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702 (2020).
- Liu, G. et al. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat. Chem. Biol.* **19**, 1342–1350 (2023).
- Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening—an overview. *Drug Discov. Today* **3**, 160–178 (1998).
- Yera, E. R., Cleves, A. E. & Jain, A. N. Chemical structural novelty: on-targets and off-targets. *J. Med. Chem.* **54**, 6771–6785 (2011).
- Wills, T. J. & Lipkus, A. H. Structural approach to assessing the innovativeness of new drugs finds accelerating rate of innovation. *ACS Med. Chem. Lett.* **11**, 2114–2119 (2020).
- Shimizu, Y. et al. AI-driven molecular generation of not-patented pharmaceutical compounds using world open patent data. *J. Cheminform* **15**, 120 (2023).
- Atanasov, A. G. et al. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
- Wright, G. D. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat. Rev. Microbiol.* **5**, 175–186 (2007).
- Ji, Y. et al. DrugOOD: out-of-distribution dataset curator and benchmark for AI-aided drug discovery—a focus on affinity prediction problems with noise annotations. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 37 8023–8031 (AAAI, 2023).
- Dias, A. L., Bustillo, L. & Rodrigues, T. Limitations of representation learning in small molecule property prediction. *Nat. Commun.* **14**, 6394 (2023).
- Tossou, P., Wognum, C., Craig, M., Mary, H. & Noutahi, E. Real-world molecular out-of-distribution: specification and investigation. *J. Chem. Inf. Model.* **64**, 697–711 (2024).
- Martinez-Mayorga, K. et al. The pursuit of accurate predictive models of the bioactivity of small molecules. *Chem. Sci.* **15**, 1938–1952 (2024).
- Grygorenko, O. O. et al. Generating multibillion chemical space of readily accessible screening compounds. *iScience* **23**, 101681 (2020).
- Nigam, A. et al. Assigning confidence to molecular property prediction. *Expert Opin. Drug Discov.* **16**, 1009–1023 (2021).
- Nada, H., Meanwell, N. A. & Gabr, M. T. Virtual screening: hope, hype, and the fine line in between. *Expert. Opin. Drug Discov.* **20**, 145–162 (2025).
- Hanser, T., Barber, C., Marchaland, J. F. & Werner, S. Applicability domain: towards a more formal definition. *SAR QSAR Environ. Res.* **27**, 865–881 (2016).
- Mathea, M., Klingspohn, W. & Baumann, K. Chemoinformatic classification methods and their applicability domain. *Mol. Inform.* **35**, 160–180 (2016).
- Wang, Z. & Chen, J. Applicability domain characterization for machine learning QSAR models. In *Proc. Machine Learning and Deep Learning in Computational Toxicology* (ed. Hong, H.) 323–353 (Springer, 2023).
- Wang, D. et al. Learning with uncertainty to accelerate the discovery of histone lysine-specific demethylase 1A (KDM1A/ LSD1) Inhibitors. *Brief. Bioinform.* **24**, bbac592 (2023).
- Soleimany, A. P. et al. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).
- Ovadia, Y. et al. Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift. In *Proc. 33rd Conference on Neural Information Processing Systems* (Curran Associates, 2019).
- Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P. & Green, W. H. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Inf. Model.* **60**, 2697–2717 (2020).

26. Grisoni, F. et al. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Sci. Adv.* **7**, eabg3338 (2021).
27. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
28. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inform.* **37**, 1700153 (2018).
29. Merk, D., Grisoni, F., Friedrich, L. & Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **1**, 1–9 (2018).
30. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
31. Iovanac, N. C. & Savoie, B. M. Simpler is better: how linear prediction tasks improve transfer learning in chemical autoencoders. *J. Phys. Chem. A* **124**, 3679–3685 (2020).
32. Kingma, D. P., Mohamed, S., Rezende, D. J. & Welling, M. Semi-supervised learning with deep generative models. *Adv. Neural Inf. Process. Syst.* **27**, 3581–3589 (2014).
33. Doersch, C. & Zisserman, A. Multi-task self-supervised visual learning. In *Proc. IEEE International Conference on Computer Vision (IEEE, 2017)*.
34. Assran, M. et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 15619–15629 (IEEE, 2023)*.
35. Kim, D., Yoo, Y., Park, S., Kim, J. & Lee, J. SelfReg: self-supervised contrastive regularization for domain generalization. In *Proc. IEEE/CVF International Conference on Computer Vision 9619–9628 (IEEE, 2021)*.
36. Sun, Y. et al. Test-time training with self-supervision for generalization under distribution shifts. In *Proc. 37th International Conference on Machine Learning 9229–9248 (PMLR, 2020)*.
37. Albuquerque, I., Naik, N., Li, J., Keskar, N. & Socher, R. Improving out-of-distribution generalization via multi-task self-supervised pretraining. Preprint at <https://arxiv.org/abs/2003.13525> (2020).
38. Pimentel, M. A. F., Clifton, D. A., Clifton, L. & Tarassenko, L. A review of novelty detection. *Signal Process.* **99**, 215–249 (2014).
39. Zhou, Y. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 7379–7387 (IEEE, 2022)*.
40. Chalapathy, R. & Chawla, S. Deep learning for anomaly detection: a survey. Preprint at <https://arxiv.org/abs/1901.03407> (2019).
41. Ruff, L. et al. Deep semi-anomaly detection. In *Proc. International Conference on Learning Representations (OpenReview.net, 2020)*.
42. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
43. Jastrzębski, S., Leśniak, D. & Czarnecki, W. M. Learning to SMILE(S). In *Proc. International Conference on Learning Representations Workshop (Curran Associates, 2016)*.
44. Özçelik, R. & Grisoni, F. A hitchhiker’s guide to deep chemical language processing for bioactivity prediction. *Digit. Discov.* **4**, 316–325 (2025).
45. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
46. Skinnider, M. A., Stacey, R. G., Wishart, D. S. & Foster, L. J. Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.* **3**, 759–770 (2021).
47. Pearce, T., Leibfried, F. & Brintrup, A. Uncertainty in neural networks: approximately Bayesian ensembling. In *Proc. 23rd International Conference on Artificial Intelligence and Statistics 234–244 (PMLR, 2020)*.
48. van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
49. Tran-Nguyen, V.-K., Jacquemard, C. & Rognan, D. LIT-PCBA: an unbiased data set for machine learning and virtual screening. *J. Chem. Inf. Model.* **60**, 4263–4273 (2020).
50. Hansen, K. et al. Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **49**, 2077–2081 (2009).
51. Manelfi, C. et al. ‘Molecular anatomy’: a new multi-dimensional hierarchical scaffold analysis tool. *J. Cheminform* **13**, 54 (2021).
52. Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **57**, 3186–3204 (2014).
53. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
54. Cereto-Massagué, A. et al. Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
55. Rossen, L., Sirockin, F., Schneider, N. & Grisoni, F. Scaffold hopping with generative reinforcement learning. *J. Chem. Inf. Model.* **65**, 6513–6525 (2025).
56. Reutlinger, M. et al. Chemically Advanced Template Search (CATS) for scaffold-hopping and prospective target prediction for ‘orphan’ molecules. *Mol. Inform.* **32**, 133–138 (2013).
57. Fooladi, H., Vu, T. N. L., Mathea, M. & Kirchmair, J. Evaluating machine learning models for molecular property prediction: performance and robustness on out-of-distribution data. *J. Chem. Inf. Model.* **65**, 9871–9891 (2025).
58. Lee, K., Lee, K., Lee, H. & Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Adv. Neural Inf. Process. Syst.* **31**, 7167–7177 (2018).
59. Boix-Adsera, E., Lawrence, H., Stepaniants, G. & Rigollet, P. GULP: a prediction-based metric between representations. *Adv. Neural Inf. Process. Syst.* **35**, 7115–7127 (2022).
60. Asinex screening library. *Asinex* [https://www.asinex.com/screening-libraries-\(all-libraries\)](https://www.asinex.com/screening-libraries-(all-libraries)) (2025).
61. Specs screening library. *Specs* <https://www.specs.net>, (2025).
62. Enamine Hit Locator. *Enamine* <https://enamine.net/compound-libraries/diversity-libraries/hit-locator-library-460>, (2025).
63. Welsch, M. E., Snyder, S. A. & Stockwell, B. R. Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.* **14**, 347–361 (2010).
64. Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors: a 2024 update. *Pharmacol. Res.* **200**, 107059 (2024).
65. Wu, P., Nielsen, T. E. & Clausen, M. H. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* **36**, 422–439 (2015).
66. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
67. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
68. Gunantara, N. A review of multi-objective optimization: methods and its applications. *Cogent Eng.* **5**, 1502242 (2018).
69. Liao, G. et al. Identification of small-molecule inhibitors of human inositol hexakisphosphate kinases by high-throughput screening. *ACS Pharmacol. Transl. Sci.* **4**, 780–789 (2021).
70. Imamura, R. M. et al. Inexpensive high-throughput screening of kinase inhibitors using one-step enzyme-coupled fluorescence assay for ADP detection. *SLAS Discov.* **24**, 284–294 (2019).
71. Puhl-Rubio, A. C. et al. Use of protein kinase-focused compound libraries for the discovery of new inositol phosphate kinase inhibitors. *SLAS Discov.* **23**, 982–988 (2018).

72. Mezna, M. et al. Development of a high-throughput screening method for LIM kinase 1 using a luciferase-based assay of ATP consumption. *SLAS Discov.* **17**, 460–468 (2012).
73. Yoshizawa, T. et al. A data-driven generative strategy to avoid reward hacking in multi-objective molecular design. *Nat. Commun.* **16**, 2409 (2025).
74. van Tilborg, D. & Grisoni, F. Traversing chemical space with active deep learning for low-data drug discovery. *Nat. Comput. Sci.* **4**, 786–796 (2024).
75. Schmidt, S., Schenk, L., Schwinn, L. & Günnemann, S. Joint out-of-distribution filtering and data discovery active learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 25677–25687 (IEEE, 2025).
76. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
77. Landrum, G. RDKit: Open-Source Cheminformatics. *RDKit* <https://www.rdkit.org> (2006).
78. Carles, F., Bourg, S., Meyer, C. & Bonnet, P. PKIDB: a curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules* **23**, 908 (2018).
79. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
80. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a ‘kneedle’ in a haystack: detecting knee points in system behavior. In *Proc. 31st International Conference on Distributed Computing Systems Workshops* (IEEE, 2011).
81. Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **103**, 3599–3601 (1981).
82. Böttcher, T. An additive definition of molecular complexity. *J. Chem. Inf. Model.* **56**, 462–470 (2016).
83. Bonchev, D., Mekenyan, O. V. & Trinajstić, N. Isomer discrimination by topological information approach. *J. Comput. Chem.* **2**, 127–148 (1981).
84. Kipf, T. N. & Welling, M. Variational graph auto-encoders. In *Proc. NeurIPS Workshop on Bayesian Deep Learning* (Curran Associates, 2016).
85. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).
86. van Tilborg, D., Rossen, L. & Grisoni, F. Zenodo repository (data). *Zenodo* <https://doi.org/10.5281/zenodo.14865513> (2026).
87. van Tilborg, D., Rossen, L. & Grisoni, F. Data for figure 2–4 for the paper “Molecular deep learning at the edge of chemical space”. *figshare* <https://doi.org/10.6084/m9.figshare.30665201.v1> (2025).
88. van Tilborg, D., Rossen, L. & Grisoni, F. Zenodo repository (code). *Zenodo* <https://doi.org/10.5281/zenodo.18846066> (2026).

Acknowledgements

This research was funded by the European Union (ERC, ReMINDER, 101077879, to F.G.). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

We acknowledge support from the Federation of European Biochemical Societies (FEBS) Excellence Award (to F.G.), the Irene Curie Fellowship, the Centre for Living Technologies and SURF (NWO grant EINF-5379 to D.v.T.). We thank A. Gardin for brainstorming on figures, R. Özçelik for technical feedback, G. Landrum for fruitful scientific discussions, S. Sueron for suggestions on the experimental work and E. R. Starr for technical feedback and proofreading the paper.

Author contributions

Conceptualization: D.v.T. and F.G.; methodology: D.v.T. and F.G.; data curation: D.v.T.; computational experiments: D.v.T.; biochemical experiments: D.v.T.; formal analysis: D.v.T.; software: D.v.T. with contributions from L.R.; visualization: D.v.T.; writing—original draft: D.v.T.; writing—review and editing: all authors.

Competing interests

The authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-026-01216-w>.

Correspondence and requests for materials should be addressed to Francesca Grisoni.

Peer review information *Nature Machine Intelligence* thanks Nessa Carson, Haicang Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted <i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All datasets and processed results are available on Zenodo at: <https://doi.org/10.5281/zenodo.14865513>. Additionally, source data and code to replicate our main

figures can be found at: <https://doi.org/10.6084/m9.figshare.30665201.v1>. All data we used from external sources is explicitly listed in the Methods section together with relevant references and links.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

| | |
|--|---|
| Reporting on sex and gender | <input type="text" value="This research does not involve human participants."/> |
| Reporting on race, ethnicity, or other socially relevant groupings | <input type="text" value="This research does not involve human participants."/> |
| Population characteristics | <input type="text" value="This research does not involve human participants."/> |
| Recruitment | <input type="text" value="This research does not involve human participants."/> |
| Ethics oversight | <input type="text" value="This research does not involve human participants."/> |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|-----------------|---|
| Sample size | <input type="text" value="Per machine learning conventions, sample sizes were determined by taking a percentage of data points from the full datasets (50% for training sets and 25% for two held-out sets). Detailed sample size information is presented in supplementary table S1. For the proof-of-concept experimental work, a sample size of 10 molecules (60 in total over 2 biological targets and 3 groups) was chosen for budget reasons."/> |
| Data exclusions | <input type="text" value="Data was excluded when molecular structures could not be processed according to our pre-processing steps as described in detail in the Methods section."/> |
| Replication | <input type="text" value="All computational experiments were replicated 10 times with different random seeds."/> |
| Randomization | <input type="text" value="Allocation of molecules in different groups was determined by their structural characteristics. Molecular datasets were divided into in-distribution and out-of-distribution subsets. Spectral clustering was first applied to molecular cyclic scaffolds (ring systems stripped of exocyclic substituents) to group structurally related compounds together. Clusters farthest apart in each dataset, corresponding to roughly 25% of all molecules, were designated as the out-of-distribution test set. The remaining compounds were further divided into an in-distribution test set (~25%) and a training set (~50%). To verify that the out-of-distribution molecules indeed originated from distinct regions of chemical space, we evaluated similarity using three complementary measures: (1) Extended connectivity fingerprints (ECFPs) calculated on cyclic scaffolds, with pairwise similarity expressed through the Tanimoto coefficient. (2) Structural core overlap, quantified as the maximum common substructure (MCS) fraction between molecular graphs. (3) Pharmacophore similarity, determined from Chemically Advanced Template Search (CATS) descriptors using cosine similarity."/> |
| Blinding | <input type="text" value="Blinding was not possible as data was split using heuristics that were engineered by the authors. However, the methods used were designed to be as unbiased as possible and do not require the tuning of any parameters (see above)."/> |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

| n/a | Involvement in the study |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Plants |

Methods

| n/a | Involvement in the study |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.