

Clever Hans in Chemistry: Chemist Style Signals Confound Activity Prediction on Public Benchmarks

Andrew D. Blevins and Ian K. Quigley

Abstract

Can machine learning models identify which chemist made a molecule from structure alone? If so, models trained on literature data may exploit chemist intent rather than learning causal structure-activity relationships. We test this by linking ChEMBL assays to publication authors and training a 1,815-class classifier to predict authors from molecular fingerprints, achieving 60% top-5 accuracy under scaffold-based splitting. We then train an activity model that receives only a protein identifier and an author-probability vector derived from structure, with no direct access to molecular descriptors. This author-only model achieves predictive power comparable to a simple baseline that has access to structure. This reveals a “Clever Hans” failure mode: models can predict bioactivity largely by inferring chemist goals and favorite targets without requiring a lab-independent understanding of chemistry. We analyze the sources of this leakage, propose author-disjoint splits, and recommend dataset practices to decouple chemist intent from biological outcomes.

1 Introduction

In a meeting years ago, an experienced chemist glanced at a panel of compounds in front of us and said “that’s a Stuart Schreiber molecule.” The observation has haunted us: if humans can recognize chemist style from structure alone, could our machine learning models be doing the same? And if so, are we training models that learn chemistry and biology, or models that simply exploit intent leakage—the stylistic regularities that arise as chemists pursue specific goals?

This failure mode is well documented in computer vision, where it goes by names like “Clever Hans” or shortcut learning [9, 20]. Models latch onto dataset signatures and spurious correlations that work in aggregate but encode the wrong mechanism [33], resulting in impressive benchmark numbers that collapse once out of distribution. We show that analogous shortcut behavior appears in medicinal chemistry.

We test this hypothesis through a two-stage experiment (Figure 1). First, we link ChEMBL assays to publication authors, creating an author–molecule graph spanning 1,815 prolific authors. We train a classifier to predict *who* made a molecule from structure alone and find that chemist styles are highly distinctive and readily learnable. Second, we train an activity model that receives only a protein identifier and the predicted author probabilities, with no direct access to molecular structure. This author-only model achieves competitive performance (Table 2), comparable to a simple ECFP+protein baseline, showing that activity can be predicted from chemist intent proxies even when direct structural descriptors are withheld.

In this work we use the term *chemist style* to refer to regularities in how molecules are designed within a lab or author group: preferred scaffolds, functional groups, reaction motifs, and target families. Operationally, however, we do not observe individual molecule designers. Instead, we infer “style” from ChEMBL document authorship metadata, treating each author identifier as a proxy for a lab- or PI-level group. This authorship-derived notion is imperfect—papers may list many authors, some of whom are biologists or non-synthetic contributors—but it is sufficient to ask whether authorship-linked patterns are strong enough to confound activity prediction.

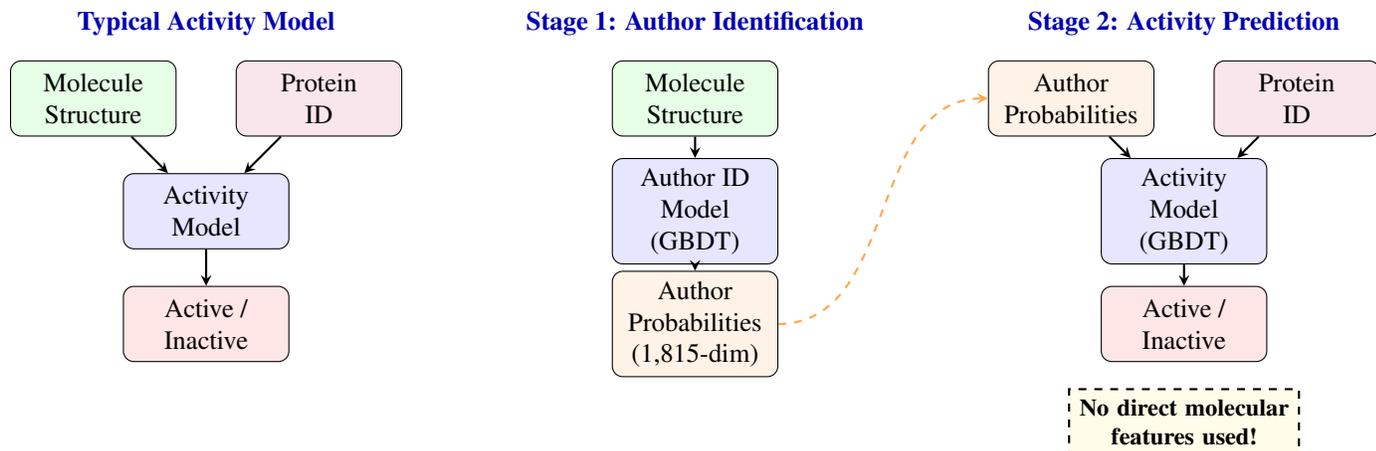


Figure 1: Overview of the two-stage chemist style leakage test. **Stage 1:** We train a model to predict which of 1,815 authors synthesized a molecule from its structure alone, achieving 60% top-5 accuracy. **Stage 2:** We use these author probabilities (plus protein ID) to predict activity, *without* providing molecular features. The strong performance (validation AUROC around 0.65) reveals a shortcut: models can predict activity by inferring chemist intent without explicitly modeling structure–activity relationships.

The problem runs deeper than known curation pitfalls [3, 8]. Community datasets aggregate heterogeneous literature under uneven experimental intent, and even rigorous standardization efforts [6, 32] leave residual signals that modern models readily exploit. We show that dataset construction choices encode chemist style signals that confound evaluation, analogous to known shortcuts in computer vision benchmarks, and we point toward leakage-aware splits, adversarial debiasing, and cross-lab replication as mitigations.

2 Related Work

A large body of protein-ligand interaction work trains and/or validates on ChEMBL-derived bioactivity or related resources (e.g., BindingDB, KIBA): examples include DeepDTA, GraphDTA, MolTrans, and DeepAffinity [13, 14, 23, 24]. Earlier ligand-based target prediction on ChEMBL demonstrated strong performance from classical ML [16, 21], and the ChEMBL resource itself continues to evolve toward higher-quality deposition and tooling [4]. Recent curation efforts (e.g. ExCAPE-DB, Papyrus, MF-PCBA) provide large, cleaned, binarized activity sets meant to reduce assay noise and ease fair benchmarking [5, 17, 32]. We situate our work within this lineage but show that author/style signals remain a strong confound.

Structure-first models now predict complex biomolecular assemblies and sometimes claim binding improvements. AlphaFold 3 introduced a diffusion architecture that models proteins, nucleic acids, ligands, and modifications in a unified framework [1]. Boltz-2 extends this line by modeling both complex structures and (approximate) binding affinities as a joint task [25]. In molecular property prediction, foundation pretraining on massive unlabeled chemical corpora (e.g., Recursion’s MolE) can yield strong ADMET and activity baselines [22]. Our results complement these advances by isolating a distinct failure mode: chemist style shortcuts that may inflate performance irrespective of the underlying biological mechanism.

Extended-Connectivity Fingerprints (ECFPs) [28] combined with boosted trees or random forests remain competitive for many ChEMBL-style tasks, particularly under leakage-prone splits [12, 26, 27, 36]. We use a gradient boosting model on ECFPs as a deliberately simple probe to measure the strength of chemist style signals and to emphasize how non-causal shortcuts can dominate performance in public benchmarks.

Computer vision has long documented dataset-specific signals and cross-dataset failures, with a promi-

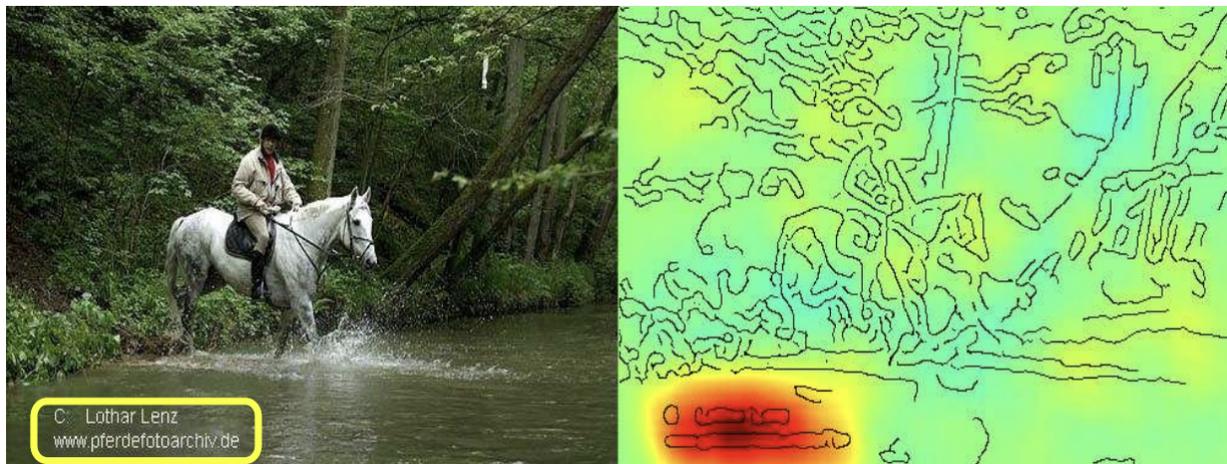


Figure 2: Classic example of a Clever Hans predictor, reproduced from Lapuschkin et al. [20]. The left panel shows an image correctly classified as containing a horse; the right panel shows a relevance heatmap revealing that the classifier bases its decision primarily on the photographer’s corner watermark rather than on the horse itself. By analogy, in this work we argue that molecules in ChEMBL carry chemist-specific “watermarks” in their structure, which even simple models can exploit, confounding attempts to learn genuine causes of binding

nent example being Torralba and Efros [33]. Subsequent work formalized *shortcuts*—decision rules that exploit unintended cues and fail under distribution shift [9]—and exposed failure modes via explanation methods [20]. Robustness benchmarks such as ImageNet-C/A [10, 11] and the WILDS benchmark for real distribution shifts [15] have driven better evaluations alongside algorithmic responses (e.g., GroupDRO and IRM) [2, 29]. Together, these literatures motivate our focus on chemist style leakage as a chemistry-specific shortcut.

Beyond well-known scaffold leakage and congeneric series effects, several works have documented systematic issues in public chemoinformatics benchmarks, including decoy and analogue bias in structure-based virtual screening (DUD-E) [7, 35], target/assay heterogeneity and label mixing [19, 31], and performance collapses around activity cliffs [34]. Methodologically, temporal or simulated-temporal splits can better approximate prospective utility than random or naive scaffold splits [18, 30]. We add *authorship* as a concrete, measurable axis of leakage with direct consequences for activity prediction, and we propose author-aware splits and diagnostics as mitigations.

3 Data

We construct an author–molecule view of ChEMBL by joining assay/compound records to publication metadata and author lists. Each molecule inherits the (multi)set of authors of the linked paper. To obtain well-supported supervision, we restrict to prolific authors (more than 30 publications and more than 600 molecules observed), yielding 1,815 authors presumed to correspond primarily to PI-level chemists or assay leads.

We study two complementary prediction problems:

1. **Author identification (1,815–way)**. Input: a small molecule; Output: a probability distribution over authors. Intended to measure the distinctiveness of “chemist style.”

2. **Activity prediction (binary).** Input: a protein identifier and the *author-probability vector* for the molecule produced by Task 1; Output: active vs. inactive. No direct molecular descriptors are provided to this model.

Task 1 (author identification). We standardize structures, deduplicate, and compute circular fingerprints to represent molecules. Labels are multi-author where applicable; evaluation counts a prediction as correct if the predicted author appears among the paper’s author set (top- k metrics analogously). Implementation specifics (fingerprint parameters, canonicalization, multi-author handling) are in App. A.1.

Task 2 (activity prediction). Starting from ChEMBL bioactivity records linked to the 1,815 authors above, we clean, normalize, and binarize activities via a Papyrus-style protocol, then build examples as compound–target pairs. Features are limited to (i) the Task 1 author-probability vector and (ii) a tokenized protein identifier; no chemical structure is exposed. Curation and featurization details are in App. A.2.

For Task 1 we perform a *scaffold split* using Bemis–Murcko scaffolds to reduce near-duplicate leakage. For Task 2 we re-join Task 1 predictions to compound–target pairs and apply a split disjoint by scaffold at the compound level.

4 Methods

4.1 Author Identification from Structure

For the author identification task, we use a standard fingerprint + gradient boosting setup.

Each molecule is standardized (canonical SMILES, salt/solvent removal) and featurized with Morgan/ECFP fingerprints (radius 2, 2048 bits, chirality on). We restrict labels to the 1,815 prolific authors described in Sec. 3 and treat author prediction as a multiclass problem. A single LightGBM classifier is trained on ECFP features to predict an author distribution over these 1,815 classes, using scaffold-grouped splits for train/validation/test. Full preprocessing and hyperparameters are given in App. A.1 and App. B.

In addition to the multiclass model, we fit one-vs-rest (OvR) binary probes—one per author—on the same features and splits. These OvR models are used only for analysis (per-author ROC–AUC / average precision distributions) and do not feed into downstream tasks.

For each compound x , we retain the multiclass softmax vector $a(x) \in \mathbb{R}^{1815}$ as an *chemist style prior*. This is the only compound-level representation used by the downstream activity model.

4.2 Activity Prediction from Author Style

The activity model operates on chemist style priors and protein identity, without access to molecular structure.

Given a compound x and target protein t , we construct an input feature vector by concatenating (i) the author-probability vector $a(x) \in \mathbb{R}^{1815}$ from the Task 1 classifier and (ii) a fixed-dimensional hashed representation of the target identifier. We use a standard feature hasher to map the string target ID to a 1,024-dimensional vector, yielding a 2,839-dimensional input.

Training labels are binary activities derived from ChEMBL using a Papyrus-style curation and binarization pipeline (App. A.2). We enforce scaffold-disjoint splits at the compound level and train over multiple random GroupShuffleSplit folds, using a held-out validation fold in each run for early stopping. A LightGBM binary classifier is then trained on these concatenated author + target features. All curation details, split construction, and model hyperparameters are described in App. A.2 and App. B.

The resulting model defines $f_\phi(a(x), t) \in [0, 1]$, an estimate of the probability that compound x is active against target t based solely on chemist style priors and protein identity.

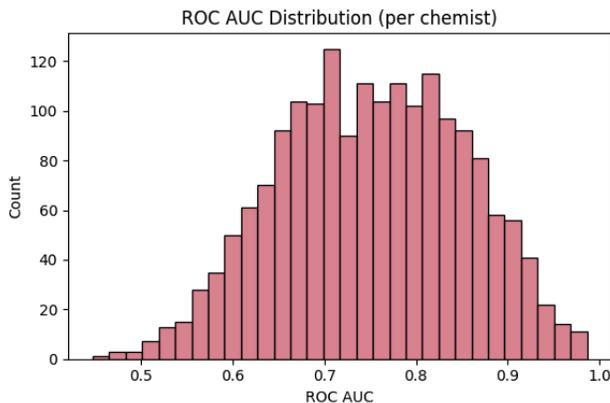


Figure 3: Histogram of ROC AUC scores of each author in one-vs-rest models.

Table 1: Author identification results on the scaffold-split held-out set.

Model	Top-1	Top-5	Top-10	Log-loss
ECFP + Gradient Boosting	0.27	0.60	0.68	3.66
Empirical class prior (dummy)	0.0003	0.0015	0.0033	7.53

5 Experiments

5.1 Author Identification Results

We first ask how well a simple fingerprint-based model can recover who made a molecule from its structure alone. Table 1 shows that the ECFP + GBDT classifier achieves nontrivial top- k accuracies on the 1,815-way author task, whereas a dummy model that only reflects empirical class frequencies is effectively at chance. In other words, even under scaffold-based splitting, chemist style is written clearly enough into molecular structure that a very simple model can reliably distinguish which lab a compound came from.

The one-vs-rest probes in Figure 3 reinforce this picture. Some authors have ROC-AUC scores close to random, but many exhibit highly separable styles: their molecules sit in distinct regions of ECFP space that are easy to carve out with shallow trees. This heterogeneity is exactly what we would expect if different labs pursue characteristic scaffolds, project types, and library designs over long periods of time. From the model’s perspective, the author label is not an incidental annotation; it is a strong, learnable signal embedded in the structures themselves.

5.2 Author-Only Activity Results

We next ask whether these author-style signals are strong enough to support activity prediction without any molecular descriptors. The author-only activity model receives only the author-probability vector $a(x)$ from Task 1 and a hashed protein identifier, but no structural features. Because the activity model only interacts with structure through this fixed author classifier, any structural detail that does not influence predicted author probabilities is effectively invisible at activity training time. In this sense, the author-probability vector is a fixed, author-centric representation of the molecule. It restricts the activity model to whatever structural information is preserved by the author classifier. We compare this to a family of baselines that all use the same LightGBM architecture but differ in which features are exposed (Table 2, Figure 4).

As expected, a random dummy model and a very simple molecular weight baseline sit close to chance.

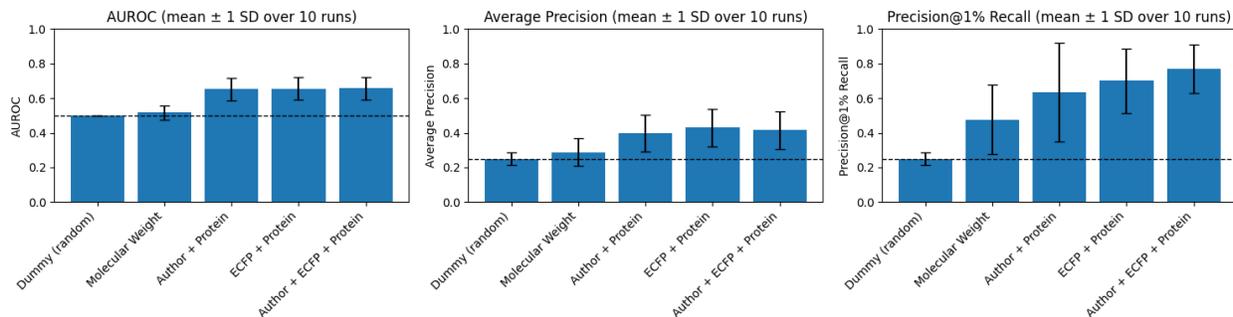


Figure 4: Validation performance of different feature sets under an identical GBDT classifier. Bars show mean performance over five random scaffold splits; error bars denote one standard deviation. The author + protein model tracks the ECFP + protein model close in AUROC, modest gap in AP. Combining author probabilities with ECFPs yields only modest additional gains, indicating that a large fraction of the predictive signal is already present in chemist style and target identity.

Table 2: Activity prediction results: mean validation AUROC, average precision (AP), and precision at 1% recall (P@1%R) over five random scaffold splits. Values are mean \pm standard deviation. Best mean in each column is **bold**.

Model / Features	AUROC	AP	P@1%R
Dummy (random)	0.500 \pm 0.000	0.249 \pm 0.036	0.249 \pm 0.036
Molecular weight + protein ID	0.517 \pm 0.043	0.288 \pm 0.079	0.477 \pm 0.202
Author probs + protein ID	0.652 \pm 0.064	0.399 \pm 0.107	0.636 \pm 0.285
ECFP + protein ID	0.656 \pm 0.067	0.430 \pm 0.107	0.700 \pm 0.188
Author probs + ECFP + protein ID	0.658 \pm 0.065	0.415 \pm 0.111	0.769 \pm 0.142

In contrast, both the author+protein and the ECFP+protein models achieve substantially higher AUROC and precision–recall performance, and—crucially—they are very similar to one another across metrics and across random splits. Adding author probabilities on top of ECFPs yields only modest additional gains. Taken together, this says that a large fraction of the predictive power we usually attribute to “learning structure–activity relationships” on this curated ChEMBL slice is also attainable by a model that never sees molecular structure at all, but only needs to understand which authors tend to work on which targets and chemotypes.

Put differently, much of the information that a simple structure-based model exploits in this setting is *explainable by chemist style*. The activity model does not need to infer detailed chemistry to perform well; it can instead learn the sociology of the dataset—how different labs behave, which series they pursue, and which targets they favor. This is precisely the Clever Hans failure mode we set out to study: strong benchmark performance that can be driven by intent leakage and lab-specific regularities and may not reflect a causal understanding of structure–activity relationships. As a consequence, we argue that good results on ChEMBL-derived benchmarks should be interpreted with caution unless author-aware splits and baselines are used to rule out this shortcut.

How structural is the author representation? As a sanity check on the representation used by the Stage 2 model, we asked how much structural information is encoded in the author-probability vectors themselves. We randomly selected a subset of hashed ECFP bits (512-bit Morgan fingerprints, radius 2) and, for each bit, trained a separate logistic-regression probe to predict the bit from the corresponding author-probability

vector on a held-out set of molecules. These simple linear decoders achieve high ROC–AUC values (with median performance around 0.9 across the probed bits) and substantial average-precision lift over the base bit frequencies, indicating that chemist-style predictions retain a rich, highly structured view of chemical space. We interpret this as evidence that public medicinal-chemistry datasets occupy a narrow “chemist-style” manifold: once a model has learned to recognize which authors a molecule most resembles, much of its circular-fingerprint representation is already determined. This reinforces our conclusion that apparent structure–activity signal on ChEMBL-derived benchmarks is tightly entangled with chemist style and data provenance.

6 Conclusion

We have shown that authorship-linked patterns in ChEMBL act as a strong shortcut for both author identification and activity prediction. A simple ECFP-based classifier can reliably recover which of 1,815 prolific authors synthesized a molecule from structure alone, and the resulting author-probability vectors, combined only with protein identifiers, support activity prediction performance that closely tracks an ECFP+protein baseline. Thus, on this benchmark slice, much of the apparent structure–activity signal can be reproduced without exposing molecular descriptors at training time, simply by learning chemist style and target portfolios.

This suggests that source identities—authors, labs, institutions, vendors, or campaigns—are an important and underappreciated confounder in public medicinal chemistry datasets. Whenever labels are drawn from many different sources, models can benefit from predicting *who* produced the data and *what* they typically work on, even if that information was never intended as an input feature.

As a result, we recommend that benchmark designers and practitioners: (i) retain and report source metadata where possible, (ii) include source-only or source+target baselines to bound the contribution of source signals, and (iii) consider source-aware splits (e.g., author-disjoint, lab-disjoint, or site-disjoint) alongside scaffold and temporal splits, particularly for headline numbers. These practices will help distinguish genuine structure–activity learning from performance that is largely driven by shortcuts on the provenance of the data.

7 Artifacts

All code, configuration files, and processed data needed to reproduce our experiments are available at github.com/Leash-Labs/clever-hans.

We will also release a leaderboard ranking all chemists in our ChEMBL subset by stylistic consistency at: <https://leash-labs.github.io/chemist-style-leaderboard>.

Acknowledgments

We thank Michael Cuccarese for inspiring this paper.

We thank the maintainers of ChEMBL and the open–source communities behind LightGBM and RDKit libraries.

References

- [1] J. Abramson, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, et al., and J. M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.
- [3] Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010.
- [4] A. P. Bento and et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 52(D1):D1198–D1206, 2024. doi: 10.1093/nar/gkad1046.
- [5] David Buterez, Jon Paul Janet, Steven J. Kiddle, and Pietro Liò. Mf-pcba: Multifidelity high-throughput screening benchmarks for drug discovery and machine learning. *Journal of Chemical Information and Modeling*, 63(9):2667–2678, 2023. doi: 10.1021/acs.jcim.2c01569.
- [6] Olivier J. M. Béquignon et al. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *Journal of Cheminformatics*, 15(1):37, 2023.
- [7] Lieyang Chen, Anthony Cruz, Steven Ramsey, Hannes Jónsson, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLOS ONE*, 14(8):e0220113, 2019. doi: 10.1371/journal.pone.0220113.
- [8] Denis Fourches, Eugene Muratov, and Alexander Tropsha. Trust, but verify: on the importance of chemical structure curation in cheminformatics and qsar modeling research. *Journal of Chemical Information and Modeling*, 50(7):1189–1204, 2010.
- [9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- [10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.
- [11] Dan Hendrycks, Kevin Zhao, Steven Basart, and et al. Natural adversarial examples. *CVPR*, pages 15262–15271, 2021.
- [12] Guang Hu, Guisheng Kuang, Wen Xiao, and et al. Performance evaluation of 2d fingerprint and 3d shape descriptors in virtual screening. *Journal of Chemical Information and Modeling*, 52(5):1103–1113, 2012. doi: 10.1021/ci300030u.
- [13] Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Moltrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021. doi: 10.1093/bioinformatics/btaa880.
- [14] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019. doi: 10.1093/bioinformatics/btz111.
- [15] Pang Wei Koh, Shiori Sagawa, and et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.
- [16] Alexandros Koutsoukas and et al. Ligand-based target prediction using ChEMBL bioactivity data and a naïve bayes classifier. *Journal of Chemical Information and Modeling*, 53(7):1957–1976, 2013. doi: 10.1021/ci300565f.

- [17] Christian Kramer and et al. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *Journal of Cheminformatics*, 14(1):82, 2022. doi: 10.1186/s13321-022-00630-4.
- [18] Greg Landrum, Martin Vogt, and Sereina Riniker. Simpd: an algorithm for generating simulated time splits on public data. *Journal of Cheminformatics*, 15(1):111, 2023. doi: 10.1186/s13321-023-00730-y.
- [19] Greg Landrum, Sereina Riniker, and et al. Combining ic_{50} or k_i values from different sources is a bad idea. *Journal of Chemical Information and Modeling*, 2024. doi: 10.1021/acs.jcim.3c01607.
- [20] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.
- [21] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Michiel Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, 2018. doi: 10.1039/C8SC00148K.
- [22] Oscar Méndez-Lucio, Christos A. Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15:9431, 2024. doi: 10.1038/s41467-024-53751-y.
- [23] Thanh Nguyen, Hang Le, and et al. Graphdta: predicting drug–target binding affinity with graph neural networks. *arXiv preprint arXiv:2005.09467*, 2020.
- [24] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. doi: 10.1093/bioinformatics/bty593.
- [25] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. preprint.
- [26] Sereina Riniker and Gregory A. Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(26), 2013. doi: 10.1186/1758-2946-5-26.
- [27] Michael C. Robinson, Robert C. Glen, and Robin D. Taylor. Reanalyzing a large-scale comparison of machine learning methods for bioactivity prediction. *Journal of Cheminformatics*, 12(1):68, 2020. doi: 10.1186/s13321-020-00457-0.
- [28] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.
- [29] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.
- [30] Robert P Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790, 2013.
- [31] Jochen Sieg, Florian Flachsenberg, and Matthias Rarey. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. *Journal of Chemical Information and Modeling*, 59(3):947–961, 2019. doi: 10.1021/acs.jcim.8b00712.

- [32] Jiangming Sun, Nina Jeliaskova, Vladimir Chupakhin, Jorge-Felix Golib-Dzib, Ola Engkvist, Lars Carlsson, Jürgen K Wegner, Hugo Ceulemans, Ivan Georgiev, and Valery Jeliaskov. Escape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of Cheminformatics*, 9(1):17, 2017.
- [33] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.
- [34] Daniël van Tilborg, Alina Alenicheva, Fabrizio Grisoni, Walters, Patrick, and et al. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23):5938–5951, 2022. doi: 10.1021/acs.jcim.2c01073.
- [35] Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of Chemical Information and Modeling*, 58(5):916–932, 2018. doi: 10.1021/acs.jcim.7b00403.
- [36] Kevin Yang, Kaleigh Swanson, Wengong Jin, and et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

A Data Cleaning and Preprocessing

A.1 Author Identification (Task 1) Preprocessing

We construct an author–molecule table from ChEMBL v33 using BigQuery. Documents (`docs_33`) are joined to activities (`activities_33`), then to compound records and structures, yielding (`doc_id`, InChI key, SMILES) pairs.

Author strings are exploded, normalized (lowercased, punctuation-stripped, whitespace-collapsed), and filtered to tokens that look like “lastname initials.” These tokens serve as author identifiers. The resulting table is deduplicated per (author, InChI key).

We then compute per-author statistics (distinct molecules and documents) and restrict to *prolific* authors with $n_{\text{molecules}} > 600$ and $n_{\text{docs}} > 30$, giving 1,815 authors. All Task 1 and Task 2 experiments are performed on molecules associated with this author set.

Molecules are standardized with RDKit (canonical SMILES, salt/solvent removal, parse failures dropped), deduplicated per (author, canonical SMILES), and assigned Bemis–Murcko scaffolds for split grouping. Features are 2048-bit Morgan/ECFP fingerprints (radius 2, chirality on). Labels are multi-author where applicable; a top- k prediction is counted as correct if any true author is in the top- k .

Scaffold-grouped splits (using Bemis–Murcko scaffolds as groups) define train/validation/test partitions. The trained multiclass model outputs a softmax vector $a(x) \in \mathbb{R}^{1815}$ for each molecule; this is cached and reused in Task 2.

A.2 Activity Prediction (Task 2) Preprocessing

For Task 2 we start from ChEMBL activities linked to the 1,815 prolific authors and apply a Papyrus-style curation:

- normalize measurements to p -scale (e.g., pIC_{50} , pK_i);
- keep only exact (non-censored) values;
- for each (compound, target, measure-type), aggregate concordant measurements (within 0.5 log units of the pair median);
- drop single-assay pairs from assays with poor reproducibility;
- binarize with an activity threshold of $pX \geq 6$.

The resulting table includes: InChI key, SMILES, target identifier, binary label, and the precomputed author-probability vector $a(x)$. Targets are represented by a 1024-dimensional feature-hashed encoding of the target ID. For models that use structure, we also compute 1024-bit ECFP fingerprints from SMILES. A simple molecular weight feature (RDKit `MolWt`) is used for a baseline.

Train/validation splits are scaffold-disjoint. The positive rate is roughly 25%, with on the order of 3×10^5 training and 8×10^4 validation examples in a typical split.

B Model and Training Details

B.1 Author Model (Task 1)

The author identification model is a LightGBM multiclass classifier trained on 2048-bit ECFP features:

- `objective = multiclass, num_class = 1815,`

- `metric = multi_logloss`,
- **typical tree parameters:** `num_leaves = 31, learning_rate = 0.05, feature_fraction = 0.8, bagging_fraction = 0.8, max_depth = 6, lambda_l2 = 1.0, is_unbalance = True`.

We train with early stopping on a scaffold-grouped validation fold, up to 2000 boosting rounds. One-vs-rest probes use the same features and splits but with a binary objective and are used only for analysis (Figure 3).

B.2 Activity Models (Task 2)

All non-dummy activity models are LightGBM binary classifiers trained on different feature combinations:

- **base parameters:** `objective = binary, n_estimators = 800, learning_rate = 0.02, num_leaves = 63, subsample = 0.8, colsample_bytree = 0.8`;
- early stopping with a patience of 50 rounds on the scaffold-disjoint validation split (monitoring AUC).

We compare:

- **Dummy:** random labels;
- **Molecular weight:** uses only the scalar molecular weight feature;
- **Author + protein:** uses $a(x)$ (1815-d) and hashed target ID (1024-d), no structural descriptors;
- **ECFP + protein:** uses 1024-d ECFP and hashed target ID;
- **Author + ECFP + protein:** concatenation of all three blocks.

For each model we report AUROC, average precision, and precision at 1% recall on the validation split (Table 2, Figure 4). A shuffled-target sanity check (not shown) confirms that performance of the author+protein model depends on genuine structure/author/target relationships rather than trivial label leakage.

C Chemist Style Examples

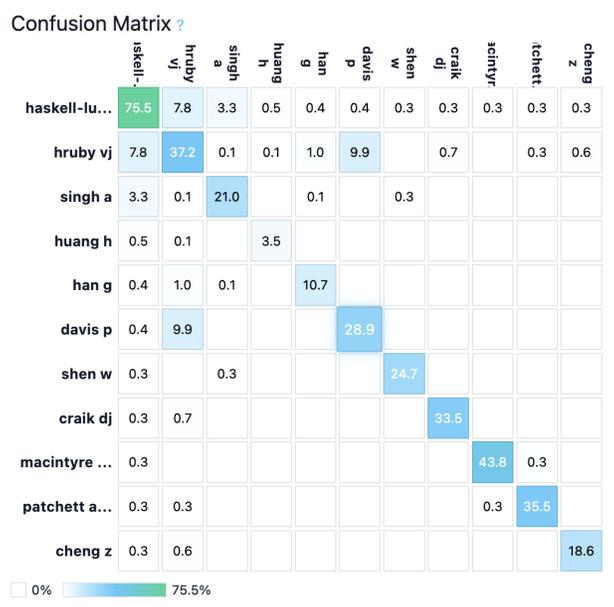
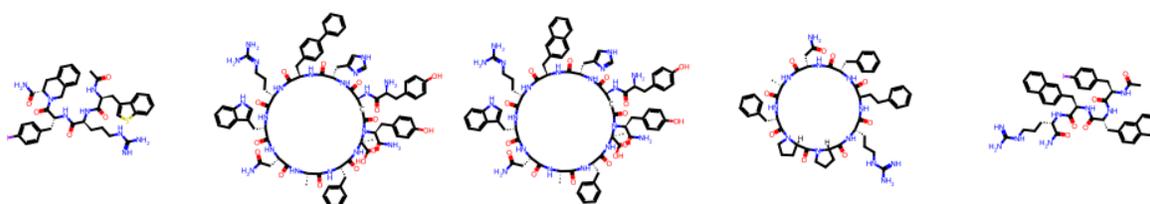


Figure 5: Five exemplar molecules from Carrie Haskell-Luevano and the corresponding author-confusion profile. The top panel shows macrocyclic melanocortin ligands built from densely functionalized, noncanonical amino acids and conformational constraints that are rare elsewhere in ChEMBL. These macrocycle-heavy series make Haskell-Luevano the most stylistically identifiable chemist in our prolific-author cohort: the author classifier assigns her as the top prediction for 75.5% of held-out molecules under scaffold splitting. The bottom panel shows the confusion-matrix row for this author, with a sharp diagonal peak and only limited spillover onto a small number of chemically similar peptide labs.

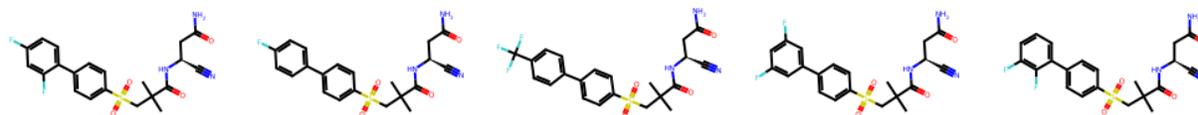
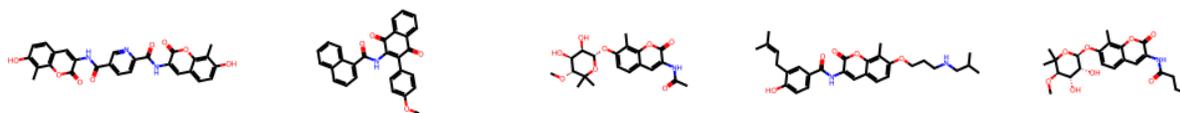


Figure 6: Representative molecules and confusion profile for T. Harrison, the third most stylistically identifiable chemist in the prolific-author set. The top panel illustrates tightly focused SAR series built around a shared, highly substituted heterocyclic or spirocyclic core with systematic variation of pendant aryl, heteroaryl, and basic side chains—the sort of NK receptor antagonist campaigns that generate many near-neighbor analogues. This combination of an unusual core scaffold and dense local SAR produces a very distinctive fingerprint signature, so the classifier assigns Harrison correctly on the vast majority of held-out molecules and otherwise confuses him with only a small cluster of nearby authors, as seen in the bottom confusion-matrix row.



Confusion Matrix ?

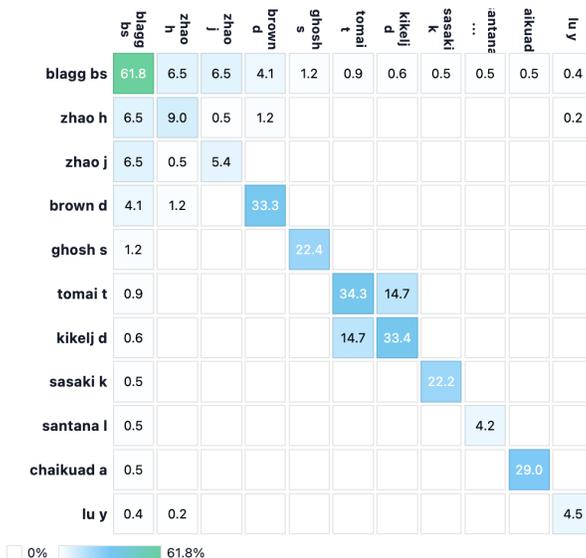


Figure 7: Representative molecules and confusion profile for Brian S. J. Blagg, ranked sixth by stylistic identifiability in our dataset. The top panel shows novobiocin-inspired Hsp90 C-terminal inhibitors and related analogues: coumarin- or benzamide-like cores coupled to heavily decorated aryl amides and noviose/noviomimetic sugar surrogates. This natural-product-derived chemotype is comparatively uncommon in ChEMBL, and the author classifier therefore predicts Blagg as the source for 68.1% of his held-out molecules from structure alone. The bottom panel shows the corresponding confusion-matrix row, where errors are concentrated on a small set of other natural-product-like series, reinforcing the idea that the model has learned a highly specific notion of “Blagg-style” chemistry.

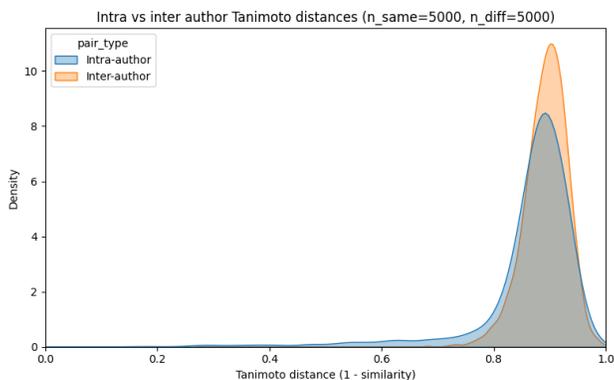


Figure 8: Plotting the Tanimoto distances between molecules that share an author vs those that don't. The ease of prediction of the author dataset cannot be explained by a simple distance metric.