[Artificial Intelligence in the Life Sciences 3 (2023) 100070](https://doi.org/10.1016/j.ailsci.2023.100070)

Contents lists available at [ScienceDirect](http://www.ScienceDirect.com/)

Artificial Intelligence in the Life Sciences

journal homepage: [www.elsevier.com/locate/ailsci](http://www.elsevier.com/locate/ailsci)

Conformal eﬃciency as a metric for comparative model assessment befitting federated learning

Wouter Heyndrickx[a](#_bookmark0),[∗](#_bookmark10), Adam Arany[b](#_bookmark1), Jaak Simm [b](#_bookmark1), Anastasia Pentina[c](#_bookmark2), Noé Sturm [d](#_bookmark3),

Lina Humbeck [e](#_bookmark4), Lewis Mervin[f](#_bookmark5), Adam Zalewski [g](#_bookmark6), Martijn Oldenhof [b](#_bookmark1), Peter Schmidtke [h](#_bookmark7), Lukas Friedrich [i](#_bookmark8), Regis Loeb [b](#_bookmark1), Arina Afanasyeva[j](#_bookmark9), Ansgar Schuffenhauer [d](#_bookmark3), Yves Moreau[b](#_bookmark1),

Hugo Ceulemans [a](#_bookmark0)

a *Janssen Pharmaceutica NV, Turnhoutseweg 30, Beerse 2340, Belgium*

b *KU Leuven, ESAT-STADIUS, Kasteelpark Arenberg 10, Heverlee 3001, Belgium*

c *Machine Learning Research, Research & Development, Pharmaceuticals, Bayer AG, Berlin 10117, Federal Republic of Germany*

d *Novartis Institutes for BioMedical Research, Novartis Campus, Basel CH-4002, Switzerland*

e *Medicinal Chemistry Department, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, Biberach an der Riss 88397, Federal Republic of Germany*

f *Molecular AI, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK*

g *Amgen Research (Munich) GmbH, Staﬀelseestraße 2, Munich 81477, Federal Republic of Germany*

h *Discngine, 79 Avenue Ledru Rollin, Paris 75012, France*

i *Global Research & Development, Merck KGaA, Frankfurter Strasse 250, Darmstadt 64293, Federal Republic of Germany*

j *Modality Informatics Group, Digital Research Solutions, Advanced Informatics & Analytics, Astellas Pharma Inc., 21, Miyukigaoka, Tsukuba-shi, Ibaraki 305-8585, Japan*

a r t i c l e i n f o a b s t r a c t

*Keywords:*

Small molecule drug discovery MELLODDY

QSAR

Federated multitask learning Applicability domain Uncertainty

In a drug discovery setting, pharmaceutical companies own substantial but confidential datasets. The MELLODDY project developed a privacy-preserving federated machine learning solution and deployed it at an unprecedented scale. Each partner built models for their own private assays that benefitted from a shared representation. Estab- lished predictive performance metrics such as AUC ROC or AUC PR are constrained to unseen labeled chemical space and cannot gage performance gains in unlabeled chemical space. Federated learning indirectly extends labeled space, but in a privacy-preserving context, a partner cannot use this label extension for performance assessment. Metrics that estimate uncertainty on a prediction can be calculated even where no label is known. Practically, the chemical space covered with predictions above an uncertainty threshold, reflects the applicabil- ity domain of a model. After establishing a link to established performance metrics, we propose the eﬃciency from the conformal prediction framework (‘conformal eﬃciency’) as a proxy to the applicability domain size. A documented extension of the applicability domain would qualify as a tangible benefit from federated learning. In interim assessments, MELLODDY partners reported a median increase in conformal eﬃciency of the federated over the single-partner model of 5.5% (with increases up to 9.7%). Subject to distributional conditions, that eﬃ- ciency increase can be directly interpreted as the expected increase in conformal i.e. low uncertainty predictions. In conclusion, we present the first indication that privacy-preserving federated machine learning across massive drug-discovery datasets from ten pharma partners indeed extends the applicability domain of property prediction models.

# Introduction

The partial virtualization of the drug discovery process is widely con- sidered as one of the more promising avenues to improve time and cost eﬃciencies in the pharmaceutical industry, including small-molecule drug-discovery modeling by means of machine learning techniques [[1]](#_bookmark26). In this context, machine learning models are typically built based on pro-

∗ Corresponding author.

*E-mail address:* [wheyndri@its.jnj.com](mailto:wheyndri@its.jnj.com) (W. Heyndrickx).

prietary and experimental compound activity datasets that are directly available to a single-partner. As the quantity and diversity of data is a central factor in the quality and usefulness of machine learning models, an appealing approach is the collaboration between competing pharma- ceutical companies and research groups through multipartner machine learning. Competing pharmaceutical companies have built up distinct chemical libraries with limited overlap [[2–4]](#_bookmark28). For instance, comparing

<https://doi.org/10.1016/j.ailsci.2023.100070>

Received 2 December 2022; Received in revised form 3 March 2023; Accepted 10 March 2023

Available online 1 April 2023

2667-3185/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

the Bayer and Schering compound libraries yielded a mere 1.5% of du- plicates, largely traceable back to external vendor chemistry, and 6.8% of shared scaffolds [[4]](#_bookmark31). A company’s library will typically drive experi- mental exploration of its own distinct chemical space, in its biology of interest. Any models trained on this experimental data will reflect such distinct chemical space. Modeling approaches that can leverage com- pound activity datasets from various companies, have the potential to also cover complementary regions of chemical space. However, com- pound libraries are at the core of the intellectual property of a pharma- ceutical research organization and as such highly confidential. There- fore, only privacy-preserving machine learning approaches can be con- sidered, which avoid directly sharing compounds, activity data, assay metadata and predictions among partners but instead leave those stored securely on the server of their respective owners behind a firewall.

The current state of collaborative multipartner learning is focused on single-task learning. A first example can be found in MAIP (Malaria Inhibitor Prediction) where multiple partners coordinated to create a single consensus model for the effect of small drug-like compounds on malaria, by exchanging the naïve Bayes classifiers and without reveal- ing any confidential training data [[5](#_bookmark33),[6](#_bookmark35)]. In another example, relating to brain tumor image segmentation, federated machine learning [[7]](#_bookmark37) was employed to leverage sensitive medical images across institutions, with- out exchanging these images [[8]](#_bookmark38). Only the model updates were sent to a secure server for aggregation and harmonization during every training round [[9]](#_bookmark39).

Multitask learning expands on single-task learning by simultaneously learning several tasks [[10](#_bookmark41),[11](#_bookmark42)]. Multitask modeling tends to boost perfor- [mance for binary classifiers in the small-molecule drug-like space [12– 19]. A beneficial effect on the predictive performance is understood to](#_bookmark44) require some degree of overlap in the compound space, in combination with correlation between the activities in this space [[16]](#_bookmark20). In a collabora- tive perspective, single-task learning implies the disclosure of an inter- est and data investment in the common task. For collaborative multitask learning on the other hand, this does not necessarily hold. Nevertheless, collaborative multitask learning, as opposed to single-task learning, is considerably less well explored in drug discovery.

The context of the current paper is the Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY) project (<https://www.melloddy.eu/>), which combines multitask learning with privacy-preserving federated learning, to unprecedentedly train on data warehouse scale datasets of ten major pharmaceutical companies (i.e., Amgen, Astellas, AstraZeneca, Bayer, Boehringer Ingelheim, GSK, Janssen, Merck, Novartis, and Servier) [[20]](#_bookmark22). In this multitask context, all partners can train their own, private tasks, unbeknownst to the other participants. During training, model updates are exchanged securely and synchronized [[21]](#_bookmark23). Thus, the model is informed by compound activities at other pharma partners, aiming for a more generalized and predictive model. To assess any such benefit, traditional performance metrics such as the area under the receiver operating characteristic curve (AUC ROC) or the area under the precision-recall curve (AUC PR) are calculated. These require known activities (labels: ‘active’/‘inactive’), hereby limit- ing their use to parts of chemical space where the evaluating partner has known activities. Additional benefit from federated learning can how- ever be expected in the chemical space of the other partners. But their compounds and activities remain hidden from the evaluating partner. Here, uncertainty metrics, which do not require knowledge of activities, offer a practical approach to evaluate such gains of federated learning. One family of uncertainty metrics originates from the spread in predictions across a family of individual models. Ensemble-based ap- proaches such as the random forest algorithm [[22]](#_bookmark24) and others [[23]](#_bookmark25) pro- duce a spread in scores across base estimators that can be consid- ered as a measure of the uncertainty. But already the single-scalar out- put from a neural network binary classifier can be interpreted proba- bilistically, often after calibration of these outputs by aligning the ex- pected error with the output by means of an independent calibration

set.

Conformal prediction (CP) is a model-agnostic framework for assess- ing the confidence in predictions [[24–27]](#_bookmark27). In the binary classification setting it states whether the prediction can be confidently assigned to one of the two classes, or not. In the latter case, assignment is to both or neither of the classes, indicating high uncertainty on the prediction. The applicability domain can be defined as the chemical space in which the model makes predictions with suﬃcient reliability [[28](#_bookmark32),[29](#_bookmark34)], imply- ing suﬃcient relevant support in the training data for the prediction. As such, the fraction of single class label predictions was proposed as a metric of the size of a model’s applicability domain. This fraction was termed as the eﬃciency from the conformal prediction framework (fur- ther referred to as conformal eﬃciency), in line with common definition [[27](#_bookmark30),[30–33](#_bookmark36)].

In summary, this work establishes the conformal eﬃciency as appli- cability domain metric and applies it to intermediate federated models from the MELLODDY project, hereby addressing the question whether such federated models have an extended applicability domain compared to local models.

# Methods

* 1. *Training data*

Each participating pharma partner committed in the range of 100,000 –2 million compounds and 350,000–35 million labels to the Year 2 MELLODDY dataset. This amounts to the bulk of their data warehouse volumes of absorption, distribution, metabolism, excretion, and toxicity (ADMET) and bioactivity data. In total, modelling in- volved more than 100,000 tasks across all partners. MELLODDY-TUNER

[[34]](#_bookmark43) is a data preparation package developed as part of the MELLODDY project. Its main features are: (i) SMILES standardization based on RD- Kit v2020.09.1.0 [[35]](#_bookmark45); (ii) binary molecular fingerprint calculation (in this work extended-connectivity fingerprints; ECFP6, folded to 32k bits)

[[36]](#_bookmark46) through the GetMorganFingerprint function in RDKit; (iii) replicate aggregation according to assay type; (iv) filtering of tasks that do not meet the task size or fold distribution quora for inclusion in training and evaluation datasets; (v) splitting of the dataset in folds according to a se- lected folding scheme, in this case based on the scaffold network [[37]](#_bookmark48) as implemented in RDKit [[38]](#_bookmark49). This approach enables different partners in the federated learning exercise to assign compounds to folds consistently without having to exchange any sensitive structural information [[39]](#_bookmark51). Five folds were created, out of which three were used for training, one for validation, and one as test fold. To avoid possible bias in evaluation, acyclic and phenylic compounds were assigned to the training set; (vi) binarization of the activity values and addition of auxiliary tasks with shifted thresholds compared to the main task of interest, while still al- lowing for user-defined expert values. From any single assay, at most five tasks were derived by varying the threshold resulting in different class balances. To avoid any dominance of assays with a higher number of derived tasks during training, MELLODDY-TUNER produces a task weighting file for further machine learning which assigns 1/N weights to all tasks, with N being the number of derived tasks from that assay [[40]](#_bookmark52). Full details on the data preparation are described elsewhere [[20]](#_bookmark22).

* 1. *Modeling*

Single-partner and multipartner model training occurred on the fed- erated platform to ensure maximal comparability and consisted of two distinct types. The so-called Phase 1 models were trained on three folds, while the Phase 2 models were trained on four folds. Phase 1 models had a validation fold available and were used to determine the optimal hyperparameters. Thereafter the Phase 2 models were trained on four folds with optimal hyperparameters from Phase 1 [[21]](#_bookmark23). Hyperparame- ter optimization for the single-partner models was performed through grid search with the parameters shown in Table S1, and by evaluat- ing AUC PR. Hyperparameter optimization for the multipartner mod-

els was through grid search with different hyperparameters, such as a higher network size to accommodate the increase of data compared to the single-partner setting. For both the single-partner and multipartner models, the same datafiles outputted by MELLODDY-TUNER were used. The network architecture was a feedforward multilayer perceptron us- ing SparseChem v0.8.2 [[41]](#_bookmark53) and minimizing the binary cross entropy as loss function through stochastic gradient descent. The implementation on the federated platform is described elsewhere [[21]](#_bookmark23). The test set was partitioned into two subsets of distinct scaffolds and comparable size, with one of the two selected as calibration set, and the other as evalu- ation set. Platt scaling [[42]](#_bookmark55) was performed in a post-processing step by fitting a logistic regression, as implemented in scikit-learn [[43]](#_bookmark56), to the logits and true labels. Predictive probabilities p(y|**x**) were transformed to predictive information entropies H by:

*𝐻* [*𝑝*(*𝑦* ***𝒙***)] = − ∑ *𝑝*(*𝑦* ***𝒙***)log2*𝑝*(*𝑦* ***𝒙***)

| | |

*𝑦*∈*𝑌*

with **x** being the input vector and y a class from the set of classes Y. Hence high entropy corresponds to high uncertainty in the class predic- tion when predictive probabilities are far from 0 to 1.

* 1. *Conformal efficiency*

In the binary classification setting, a prediction with a conformal predictor can have four distinct outcomes: a prediction set containing both class labels {0,1}, none of the class labels {}, or a single class label

{0} or {1}. The metric of conformal eﬃciency is defined as the fraction of single class label predictions:

trivially achieved by producing prediction sets containing both class la- bels {0,1} regardless of the input [[47]](#_bookmark61). Hence, a proper nonconformity function is required to achieve a reasonable eﬃciency.

Irrespective of the measure of uncertainty, it should be noted that there are two sources of uncertainty: aleatoric uncertainty arises from the inherently stochastic nature of the data (e.g., a compound with ac- tivity close to the activity threshold), while epistemic uncertainty arises from the lack of similar data [[48]](#_bookmark62). Different uncertainty measures per- tain to different aspects of uncertainty. While the spread in model out- puts from an ensemble-based model measures the epistemic uncertainty, the output itself is related to both epistemic and aleatoric uncertainty [[48](#_bookmark62),[49](#_bookmark63)]. As the nonconformity function in this work was based on the model output, the conformal eﬃciency will reflect both epistemic and aleatoric uncertainty.

The conformal eﬃciency per task is calculated as metric, and sub- sequently the difference between the multipartner and single-partner

setting was made to obtain the ΔCE ([Fig. 1](#_bookmark11)). Only tasks of suﬃcient

size, class balance, and predictive quality were considered. Concretely,

a size quorum of minimally 25 actives and inactives in the calibration set was imposed. For Phase 1 models this corresponded to the valida- tion fold, while for Phase 2 models this corresponded to approximately half of the test fold. In addition, tasks meeting this quorum but failing to achieve an AUC ROC of 0.6 or higher on the calibration set were dropped from the analysis. Only tasks of main interest were consid- ered. Any auxiliary tasks, including those resulting from thresholding

schemes, were similarly dropped. A significance level (*𝜀*) correspond-

ing to an expected maximal error rate of 0.05 and hence a confidence

level of 0.95 was selected unless stated otherwise. Higher eﬃciencies are typically observed at higher significance levels such as 0.20 [[47]](#_bookmark61).

*CE* =

*𝑛*{1} + *𝑛*{0} =

*𝑛*{1} + *𝑛*{0} + *𝑛*{ } + *𝑛*{0*,*1}

*𝑛*{1} + *𝑛*{0} N

In a practical drug discovery setting, where false positives can be costly due to subsequent unsuccessful experimental validation, lower signifi- cance levels such as 0.05 are preferable since these will typically push

With e.g. n{1} indicating the number of single active class label com- pounds, and N the total number of predicted compounds for that task.

Higher values can be a consequence of two effects: (i) having a larger separation between the two classes, lowering the instances attributed

class. A positive difference in conformal eﬃciency (ΔCE) between the to both classes, or (ii) having fewer instances not attributed to either

multipartner and single-partner model is associated with uncertainty reduction. Inductive Mondrian CP is applied, where the prediction for each class is estimated separately using an individual calibration set per class [[26](#_bookmark29),[30](#_bookmark36),[44](#_bookmark58),[45](#_bookmark59)]. The code is based on an implementation from Toccaceli [[46]](#_bookmark60).

The conformal prediction framework provides formal guarantees on error rates, by calibrating a nonconformity function, most often based on the predictive probability outputted by the model. Exchangeability is assumed between calibration and test set, which is related to but slightly weaker than the independently and identically distributed (IID) condi- tion. In situations where the test set less closely resembles the calibra- tion set, the assumption of exchangeability might not hold, and no for- mal statements on the error rates can be made. As is discussed below, such a drop in exchangeability is not sudden and abrupt, but gradual upon moving away from the calibration set, similar to a type of ero- sion or wear, and does not necessarily eliminate the usefulness of the conformal prediction framework and its quantities such as eﬃciency. An appropriate choice of nonconformity function is essential for a con- formal predictor. Ideally, the nonconformity scores are related to the uncertainty in the class prediction. Since the single-scalar output from a neural network binary classifier already is indicative of the error rate and hence the uncertainty, basing the nonconformity function on the model’s output is a natural choice. Concretely, the nonconformity func-

active class p(y=1|**x**) for the inactives and actives, respectively. It can tion here was the positive and negative predictive probability of the

be noted that any function is potentially applicable as nonconformity function. Even an uninformative constant function independent of the input can be selected, but in this case the guaranteed error rates will be

models to more conservative behavior by increasing n{0,1} and decreas- ing n{1} [[26](#_bookmark29),[44](#_bookmark58)], leading to higher confirmation rates (albeit lower than the confidence level) [[50]](#_bookmark64).

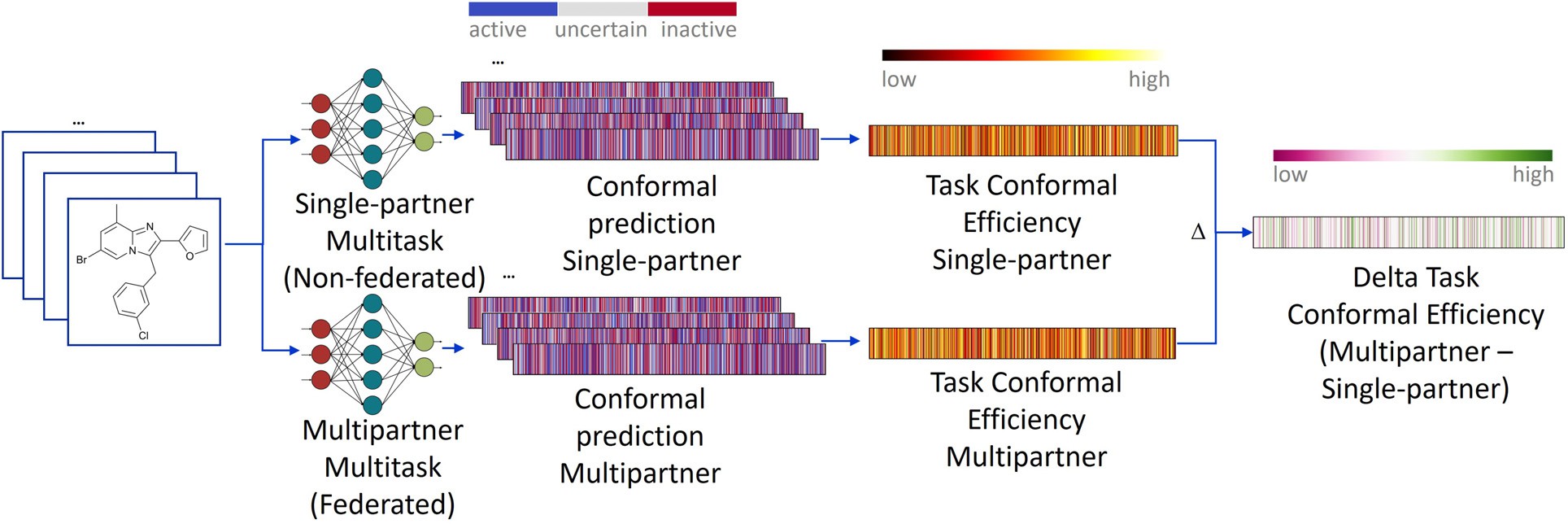
* 1. *Datasets for conformal efficiency exploration*

For exploring the relationship between the conformal eﬃciency and the predictivity, three datasets were used. The first two datasets were created by splitting the test fold in two parts (termed ‘1/2 test‘ and ‘2/2 test’) and preventing scaffolds based on the scaffold network [[37]](#_bookmark48) to occur in both parts simultaneously. A third dataset (termed ‘unseen’) was created by selecting compounds that were withheld from the MEL- LODDY dataset, and compounds that were generated after the composi- tion of the MELLODDY dataset. Only the scaffolds not already occurring in the MELLODDY dataset were retained. So, the latter dataset contains compounds that are, from a scaffold point of view, novel to the model.

* 1. *Datasets for uncertainty estimation*

[Table 1](#_bookmark12) shows an overview of the datasets used for uncertainty es- timation. Several datasets are designated as unlabeled, indicating that no activity data was available for these compounds. On the other hand, the labeled type indicates that the compound-task combinations carried labels derived from experimental measurements.

Regarding the labeled datasets, ‘Training folds 1–3′, ‘Validation fold’, and ‘Test fold’ resulted from the five-fold split of the internal MELLODDY dataset from the individual pharma partners, respectively. Only labeled compound-assay pairs were included in this dataset, allowing model training and calculation of performance metrics requiring labels. Any unlabeled data points for compounds in these folds were excluded from these datasets (but these compounds were eligible for sampling to com- pose unlabeled datasets, see below). For Phase 2 models, the validation fold was part of the training data, and the evaluation was carried out on half of the test fold, with the other half used for calibration.



**Fig. 1.** Schematic of the conformal eﬃciency metric calculation.

**Table 1**

Overview of datasets used for uncertainty estimation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Index | Size (# compounds) | Selection | Source | Type | Median number of set ECFP bits |
| Training fold 1 (labeled compound-assay pairs)§ | 1 | Large∗ | All | Internal | Labeled | Partner-dependent |
| Training fold 2 (labeled compound-assay pairs)§ | 2 | Large∗ | All | Internal | Labeled | Partner-dependent |
| Training fold 3 (labeled compound-assay pairs)§ | 3 | Large∗ | All | Internal | Labeled | Partner-dependent |
| Validation fold (labeled compound-assay pairs)§ | 4 | Large∗ | All | Internal | Labeled | Partner-dependent |
| Test fold (labeled compound-assay pairs)§ | 5 | Large∗ | All | Internal | Labeled | Partner-dependent |
| Compounds in training folds§ | 6 | 10,000 | Randomly sampled | Internal | Unlabeled | Partner-dependent |
| Compounds in validation fold§ | 7 | 10,000 | Randomly sampled | Internal | Unlabeled | Partner-dependent |
| Compounds in test fold§ | 8 | 10,000 | Randomly sampled | Internal | Unlabeled | Partner-dependent |
| ExCape-ML[19] | 9 | 10,000 | Randomly sampled | External | Unlabeled | 68 |
| ChEMBL 25 - MELLODDY-TUNER[34] | 10 | 10,000 | Randomly sampled | External | Unlabeled | 69 |
| DUD-E[54] | 11 | 10,000 | Randomly sampled | External | Unlabeled | 68 |
| SCUBIDOO[57] | 12 | 9994 | All | External | Unlabeled | 63 |
| Commercial catalog 1‡ | 13 | 5309 | All | External | Unlabeled | 61 |
| Commercial catalog 2‡ | 14 | 10,000 | Randomly sampled | External | Unlabeled | 64 |
| DrugSpaceX[58] | 15 | 10,000 | Randomly sampled | External | Unlabeled | 78 |
| FDA-approved small-molecule drugs[55,56] | 16 | 3981 | All | External | Unlabeled | 53 |

‡ Source is undisclosed.

§ See methods section for additional details.

∗ Exact size in terms of number of compounds differs between the partners but substantially exceeds 10,000 in all cases.

Regarding the unlabeled datasets, a variety of external and internal datasets was assimilated in order to evaluate the models’ uncertainty, as shown in [Table 1](#_bookmark12). For each of these unlabeled datasets, a 10,000- molecule sub-sample was selected, unless there were less molecules available, in which case no sampling was done, and the full dataset se- lected. The collection of datasets aimed to cover a wide chemical variety, while guaranteeing the property of drug-likeness.

The public dataset ExCape-ML [[19]](#_bookmark21) combines a number of published libraries such as PubChem [[51]](#_bookmark65) in an effort to consolidate the publicly available bioactivity data. The ChEMBL 25 [[52]](#_bookmark67) dataset was prepared by the MELLODDY consortium and released through MELLODDY-TUNER [[34]](#_bookmark43).

The DUD-E (Directory of Useful Decoys, Enhanced) dataset contains commercially available compounds from ZINC [[53]](#_bookmark68), which are suitable as decoys derived from an active ligand in docking experiments against a specific target [[54]](#_bookmark70). Here the 10,000 compounds were drawn from the decoys for the ‘diverse’ set of targets consisting of AKT1, AMPC, CP3A4, CXCR4, GCR, HIVPR, HIVRT, and KIF11. It can be mentioned that although ligand-based machine learning approaches with the DUD- E dataset would be problematic due to the decoys being selected as maximally dissimilar to the targets’ known ligands [[54]](#_bookmark70), here merely the predictions for other targets and assays are of interest.

The library labeled ‘FDA-approved small-molecule drugs’ contains all small-molecule drugs from the DrugCentral public resource [[55](#_bookmark71),[56](#_bookmark72)]. Data extraction was performed in September 2019.

The ‘Commercial catalog 1 and 2′ datasets were obtained from undis- closed commercial vendors and contain drug-like compounds typically used in early screening.

In addition to these datasets containing existing external chem- istry, an attempt was made to include novel chemistry that exists only virtually but is still synthetically tractable. SCUBIDOO is an ex-

ample of such a dataset where ∼8000 existing building blocks were

combined with 58 robust organic reactions commonly employed in

drug discovery, resulting in 21 million virtual molecules. Here the ‘S’ dataset was used, containing 9994 molecules, which constitutes a small but representative sample of the full dataset [[57]](#_bookmark73). Another ex- ample is the DrugSpaceX database which has been created by apply- ing expert-defined transformations such as bioisostere replacements, ring opening/closing, homologization or linker replacements to 2215 approved small molecule drugs, resulting in over 100 million com- pounds. The 10,000 selected compounds were sampled from the ‘sam- ple set’ (S) containing 937,230 molecules after a first transformation step [[58]](#_bookmark74).

Internal libraries were also included into the analysis, which origi- nate from each of the pharma partners. While some of these compounds are known in the public space, the majority is presumed to exclusively form part of the internal library and is subject to privacy concerns. The compounds in the training, validation and test fold ([Table 1](#_bookmark12)) were hence a set of internal compounds partitioned through the fold splitting tech- nique.

While there may be some degree of overlap with the internal MEL- LODDY datasets for these unlabeled datasets (e.g., publicly available compounds may be part of the internal library or may even belong to the training or validation set), this was not expected to contribute signif- icantly to the results due to the small overlap expected between public and proprietary datasets and the sparsity of the labeled activity ma- trix (far below 1%). Indeed, predictions on the unlabeled datasets are dense, i.e. across all tasks, while the predictions on the labeled datasets are sparse, since a prediction is only generated where there is an ac- tual label available. The degree of identical overlap of the unlabeled datasets with the internal MELLODDY datasets in the 32k bit descriptor space was investigated by three pharma partners, and the findings were in line with expectations: (i) the FDA-approved small-molecule drugs were clearly most highly overlapping with the internal library, suppos- edly due to their use as reference compounds; (ii) the virtual libraries (SCUBIDOO, DrugSpaceX) had very low overlap (but nonzero); and (iii) the other external datasets were in-between these two extremes.

# Results

* 1. *Outline*

This work addresses the question whether federated models have an extended applicability domain as measured through the conformal eﬃciency compared to local models. To this end, the following steps were undertaken:

1. The behavior of conformal eﬃciency as uncertainty metric was in- vestigated, contrasted with other uncertainty metrics, and found to behave intuitively.
2. It was shown that conformal eﬃciency on a task level, even in sce- narios most challenging the exchangeability assumption, still corre- lated with performance metrics requiring labels. Hence, it can be argued that the uncertainty from conformal eﬃciency was justified.
3. This correlation broke down under similar conditions for ΔCE, how-

ever the task-aggregated ΔCE maintained its direction.

1. The task-aggregated ΔCE was applied to the local and federated

cability domain for the latter. This result was confirmed with ΔPlatt- model and revealed lower uncertainty and hence an extended appli-

scaled entropy.

* 1. *Comparison of uncertainty metrics*

In this section, the choice for selecting the conformal eﬃciency over other, established uncertainty metrics is elaborated upon, by demon- strating some key findings, without therefore embarking on an exten- sive benchmarking study. In the spirit of transparency and reproducibil- ity, calculations were performed on public datasets with code that was made publicly available. The metrics included single model approaches, where predictive probabilities can be used directly as input for the non- conformity function of the conformal predictor or the calculation of the predictive entropy. In case of predictive entropy, the effect of calibra- tion by Platt scaling is investigated. Ensemble-type approaches consider uncertainty metrics based on the spread of multiple predictive prob- abilities, such as the variance of the predictive probabilities and the information gain (IG; also known as mutual information) [[59](#_bookmark76),[60](#_bookmark47)].

Both variance and IG are established methods to estimate the un- certainty, but in the studied setting these uncertainty metrics exhibited overconfidence. Indeed, Fig. S1 demonstrates that, counter-intuitively, the uncertainty of these metrics decreased with distance from the train- ing set. The overconfidence in predictions on out-of-domain data points of neural networks (including networks with rectified linear unit (ReLU) non-linearities) has been observed elsewhere) [[61](#_bookmark50),[62](#_bookmark51)] and deeper anal- ysis falls outside the scope of this paper. In contrast, the predictive en- tropy exhibited the expected behavior of increasing uncertainty under the same conditions, as did the conformal eﬃciency (Fig. S2).

To estimate the robustness of the uncertainty metrics, their behav- ior was investigated under changing model conditions, such as training data volume and hidden layer size, influencing the regularization of the model. A clearly overtrained model might produce predictive probabil- ities close to 0 and 1, indicating low uncertainty, while a model with excessive regularization might produce predictive probabilities close to

0.5. Fig. S3 demonstrates this behavior. By varying the size of the hid-

entropy was high at very small network sizes (*<*5) for predictions in den layer in a single-layer network for a public benchmark dataset,

both labeled and unlabeled space, reflecting higher uncertainty. It then dropped sharply with increasing layer sizes, indicating very low uncer- tainty. Regarding the training set size, larger training sets presumably extend the chemical space covered by the model, resulting in a lower uncertainty on a common test set. However, such an effect was not observed for the uncalibrated entropy. For the largest networks, only the smallest training set of 100 training samples displayed somewhat higher entropy values compared to the other training set sizes. Apply- ing calibration to the predictive probabilities through Platt scaling did effectively address this overconfidence. Also, the Platt-scaled entropy favorably did not display a strong dependency on the degree of regular- ization through the hidden layer size, and when more training data was involved, a lower uncertainty is observed, aligned with expectations.

Calibration to estimate class probabilities through the conformal pre- diction framework had a similar effect on the uncertainty estimation as calibration *via* Platt scaling. No pronounced dependency on the degree of regularization through the layer size was found and adding training data reduced uncertainty as expected through higher eﬃciency. These findings underline the beneficial effect of calibration, either through the conformal predictor framework, or Platt scaling, when comparing pre- dictive probabilities from two distinct models.

Finally, both conformal eﬃciency and the Platt-scaled entropy show favorable characteristics such as (i) a relative insensitivity to a wide range of hyperparameter selections, (ii) an expected behavior when moving away from the training set, and (iii) an expected behavior in relation with the amount of data added. In this work, the conformal ef- ficiency was selected due to its flexibility as distribution-free method and the interest conformal prediction has attracted in the industry in the context of small molecule activity modelling [[30–32](#_bookmark36),[63](#_bookmark52)].

* 1. *Conformal efficiency*

The metric of eﬃciency is defined as the fraction of single class label predictions, ({0},{1}). Such single class label predictions are consid- ered to signal confidence of the model as one of both class assignments is substantially more likely than the other. On the other hand, predic- tion sets with none ({}) or both labels ({0,1}) constitute low confidence and high uncertainty. In the current study, most of the modelled tasks generated high uncertainty prediction sets containing both labels, while empty prediction sets were absent.

Decreases in uncertainty measured by ΔCE between the multipartner

and single-partner model suggest better generalization of the multipart-

ner model to some regions outside a partner’s chemical space. However, a decrease in uncertainty, albeit a positive sign, does not necessarily im- ply an increase in predictive performance. In other words, decreases in uncertainty might not be justified. To address this point, the relationship between the conformal eﬃciency, and more established performance metrics that require labels was explored from both a theoretical and empirical perspective.

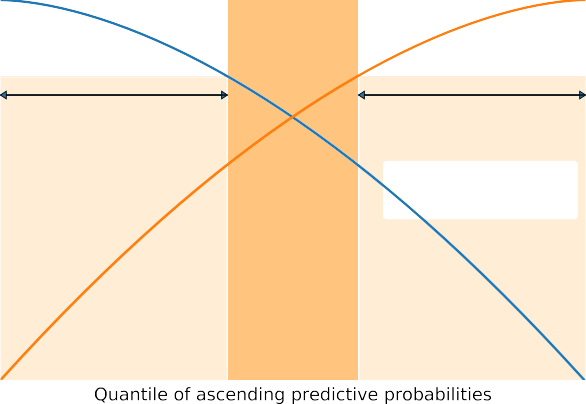
First, a theorical relationship between the conformational eﬃciency and the recall (TPR) and specificity (TNR) was established and illus- trated in [Fig. 2](#_bookmark13). This treats a single typical task, with low significance level. Three regions along the x-axis can be observed, delimited by qn

and qp. These are the quantiles associated with the thresholds for nega-

As such, 0*<*q*<*qn contains single inactive predictions {0}, and the tive and positive class label predictions, respectively, over all classes. negative eﬃciency equals the width of the area. 1*>*q*>*qp contains single

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

**Fig. 2.** Relationship between the recall (TPR) and specificity (TNR) on one hand, and the conformal eﬃciency on the other. Stylized representation of a



typical case at a low significance level (*𝜀*), for one modelled task. It should be

noted that realistic specificity and recall curves will not be smooth nor symmet- ric (see Fig. S5 for a numerical example and the provided code for generation).

qn and qp are the quantiles associated with the thresholds for negative and pos- itive class label predictions, respectively.

this area. The remaining area, qp *<*q*<*qn, contains both class label pre- active predictions {1}, and the positive eﬃciency equals the width of

negative eﬃciency and qp *<*q*<*qn does not contribute to the eﬃciency. diction sets {0,1}. The total eﬃciency is the sum of the positive and

The validity of the calibration set will be at least 1-*𝜀* by design, for both The validity is defined as the fraction of compounds labeled correctly.

classes individually in class-balanced Mondrian CP.

The dark orange area will contribute fully to the validity for both classes, since assigning both labels is always correct. The light orange ar- eas contribute to the validity for one of both classes, while the white area

contains the incorrect predictions, amounting to *𝜀*∗P or *𝜀*∗N predictions.

For the actives, the recall will be 1-*𝜀* at qp, while for the inactives, the specificity will be 1-*𝜀* at qn. This is because CP will set qp and qn to allow at most an error rate of *𝜀*, implying FN≤*𝜀*∗P and FP≤*𝜀*∗N. Hereby follows 1-(FP/N)=1-FPR=TNR≥1-*𝜀* for the negative class at qn. In other words,

the conformal predictor will refrain from assigning class labels for the riskiest samples, which puts a lower bound on the recall/specificity.

Hence for the calibration set, there is a direct link between the re- call/specificity and the eﬃciency. qn and qp will shift according to the quality of the classifier. Better class separation and accompanying

higher recall/specificity will push qn and qp closer together, hereby increasing the eﬃciency. In summary, the specificity at the quantile threshold for negative class label prediction qn for the calibration set

determines the positive eﬃciency and vice versa. A similar relationship

exists for the recall and the negative eﬃciency. Hence it follows that eﬃciency can inform on the predictive performance of the model.

In the preceding section the calibration set was considered for both calibration and eﬃciency calculation. Exchangeability in this situation was guaranteed since the datasets were identical. In practice, consid- erations of exchangeability become relevant when applying conformal prediction to unseen datasets. The strength of the link between the recall/specificity and the eﬃciency will depend on the degree of ex- changeability, and this is explored empirically in the following section. Three datasets were considered. Arguably, the ‘1/2 test’ and ‘2/2 test’ datasets were closer to the training set of the model than the ‘unseen’ dataset, since the latter contained novel scaffolds added to the internal library only after the composition of the MELLODDY dataset. As the ‘1/2 test’ and ‘2/2 test’ were part of the MELLODDY dataset, one might expect scaffolds present here to bear more similarity to the scaffolds in the training set, as they could be to a greater degree the result of incremental changes to scaffolds. The ‘unseen’ dataset was considered to

be the best approximation of true unlabeled space, i.e. novel chemistry outside of a partner’s chemical domain where no labels were available. From [Fig. 3](#_bookmark14), the impact of successive deviations from the ideal- ized case is illustrated. In an idealized case, with perfect exchange- ability, there is a clear relationship between the eﬃciency and the re-

call/specificity.

The recall and specificity both consider one point on their respective curves, corresponding to one threshold. Therefore, a first deviation that could be considered was switching recall/specificity for a predictive per- formance metric that considers the full range of thresholds. Calibrated AUC PR, a rebalanced version of AUC PR facilitating task comparisons [[64]](#_bookmark54), is such a metric. In this case, the relationship between the eﬃ- ciency and predictive performance was conserved quite well over all tasks, with a Pearson correlation coeﬃcient (r) of 0.88.

A second deviation from the idealized case was swapping either the calibration or the evaluation set. This was thought to present a chal- lenge to the exchangeability, since the datasets were constructed in such a way to prevent any scaffolds occurring in both. Two effects could be observed. The effect of changing the calibration set was somewhat more impactful than changing the evaluation set. And moving from ‘1/2 test’ to ‘unseen’ affected the r more than from ‘1/2 test’ to ‘2/2 test’, presum- ably due to the similarity and higher exchangeability between the latter two.

A final deviation consisted of swapping both the calibration and the evaluation set. The lowest r value was observed for the ‘unseen’ dataset as calibration set, which is consistent with the previous observations. Despite these conditions that maximally challenged exchangeability, still a reasonable correlation r near 0.5 was being observed. This sug- gests that at least some correlation will remain when pushing out further into unlabeled space.

Having established the stability of the relationship between the pre- dictive performance and the conformal eﬃciency, a next step was to investigate how the conformal eﬃciency itself behaved under devia- tions from perfect exchangeability. The upper half of [Fig. 4](#_bookmark16) illustrates that the predictive performance was highly correlated between ‘1/2 test’ and ‘2/2 test’. This correlation dropped when challenging exchangeabil- ity more with the ‘unseen’ dataset, with a lower predictive performance on the ‘unseen’ dataset, in part reflective of temporal data drift. This is mirrored in the conformal eﬃciencies, when calculated on the same evaluation set and varying the calibration set. Hence, calibrating the conformal predictor on the ‘unseen’ dataset led to lower eﬃciencies. The effect of varying the evaluation set and keeping the calibration set constant was similar but less pronounced, leading to higher r (Fig. S4). In the lower half of [Fig. 4](#_bookmark16) the correlation between the deltas for different datasets is shown to have dropped substantially, being much

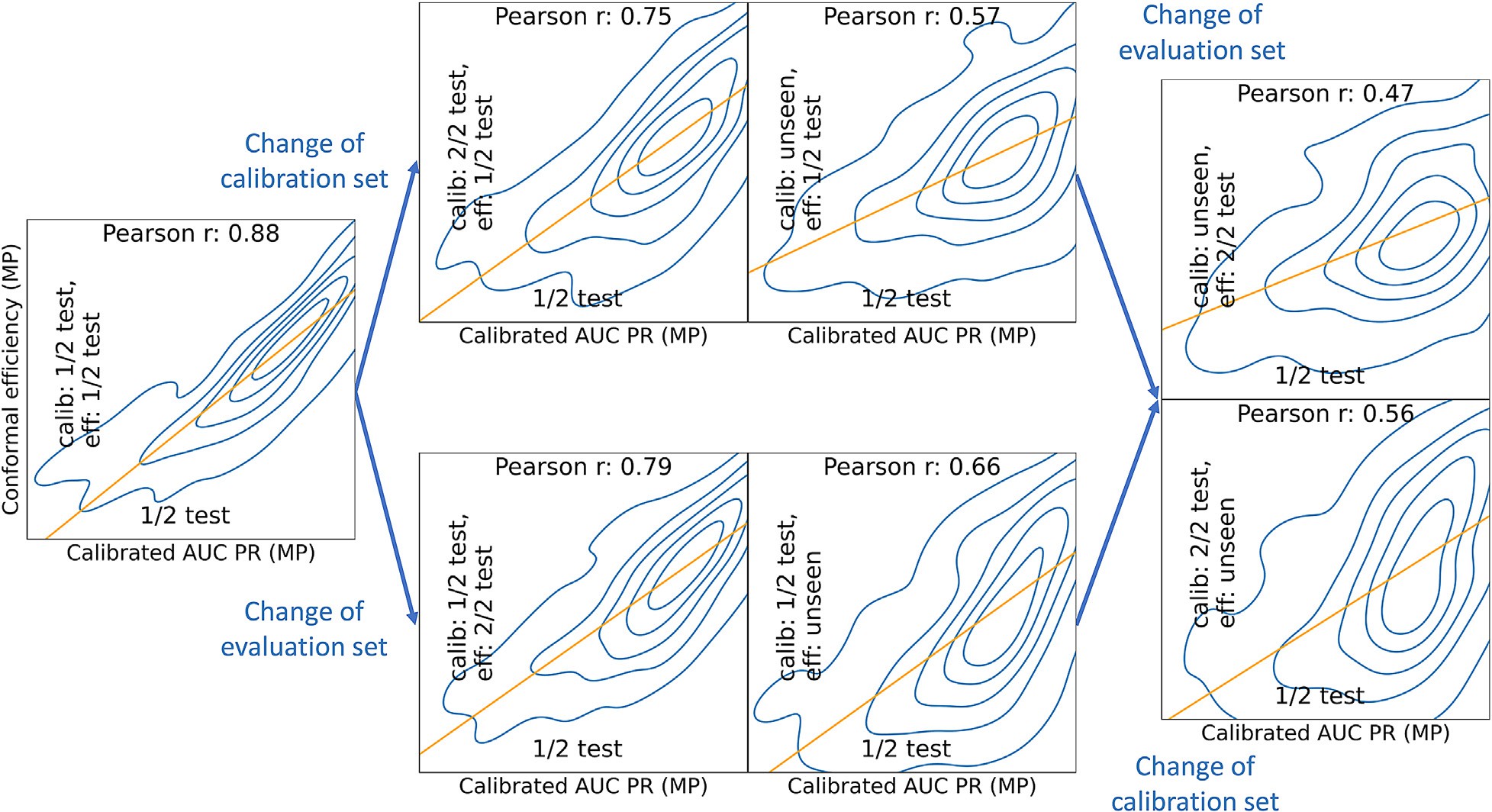
closer to zero, for both the delta calibrated AUC PR and the ΔCE. The

finding to some degree compromises a task-level performance evalua-

tion between the single-partner and multipartner models, since observed gains would not be consistent between datasets. On an aggregated level, the effect was more stable, as the highest density of tasks was routinely found in the upper right quadrant, regardless of the dataset. To ensure that the observed shift from the origin was not due to stochastic effects, a permutation analysis was performed where the set of single-partner and multipartner model predictions were randomly interchanged for a single task. So, a task had only two options: it kept it predictions as is, or it exchanged all predictions between single-partner and multipartner. Due to the multitude of tasks, many combinations were possible. It was found that the observed shift was much larger than expected based on chance if single-partner and multipartner models were producing inter- changeable predictions ([Fig. 4](#_bookmark16), bottom right). This suggested that the

retrieved from the overall ΔCE. As the gains in predictive performance overall gains in predictive performance in the multipartner case, can be

for the multipartner model are discussed in detail elsewhere [[20]](#_bookmark22), we refrain from further analysis here, but merely establish the link between the delta predictive performance and the delta conformal eﬃciency at an aggregated level.



**Fig. 3.** Linear correlation between the predictive performance (calibrated AUC PR) and the conformal eﬃciency upon successive deviations from an idealized case with perfect exchangeability. Plots show the density of tasks (kernel density estimation). The linear trendline is shown in orange. The ‘1/2 test’ dataset is consistently used for calibrated AUC PR calculation.

The ‘unseen’ dataset was considered to be the best approximation of true unlabeled space, i.e., novel chemistry outside of a partner’s chem- ical domain where no labels were available. The lack of labels prohib- ited metrics such as the calibrated AUC PR, and conformal eﬃciency became the estimate of choice when projecting out to unlabeled space. Hence, a realistic situation when evaluating the MELLODDY models for generalization to novel types of chemistry, was to calibrate on the test fold, and calculate the conformal eﬃciency on an unlabeled set, here approximated by the ‘unseen’ dataset. [Fig. 5](#_bookmark17) shows the effect of that situation against a reference where exchangeability was ensured since the ‘unseen’ dataset was used for both calibration and calculation of the eﬃciency. In this situation the link to the predictive performance was better preserved. Calibrating instead on a part of the test set now weak- ened the relationship between the task conformal eﬃciencies. For the middle and higher end of the eﬃciency range, the observed eﬃciencies tended to overestimate the reference eﬃciencies. For the aggregated val- ues, in conclusion, calculating the conformal eﬃciency on an unlabeled set, might have overestimated the predictive performance by roughly 25% based on the location of the maximal density. This overestimation can be related to the reported drop in eﬃciency upon updating the cal- ibration set without model retraining to account for data drift [[33](#_bookmark40),[65](#_bookmark55)]. Indeed, generally higher eﬃciencies were observed for predictions on the ‘unseen’ dataset with part of the test fold as calibration set (reflective of data drift), than with the ‘unseen’ dataset as calibration set (absence of data drift).

* 1. *Federated gains*

The aggregation of the ΔCE for all partners over all unlabeled datasets is shown in [Table 2](#_bookmark15), revealing that all partners showed a pos-

itive median ΔCE. Lower uncertainty predictions in the multipartner

setting appear as positive values, and vice versa, any increase in uncer-

tainty is reflected in negative values. The median increase over all part- ners was 5.5%, while the highest reported increase for a partner was

ported 0% to 5% median ΔCE gains regardless of the type of chemistry 9.7%. Further zooming in on the individual partners in [Table 2](#_bookmark15), half re-

a 10% gain in ΔCE corresponded to the classifier producing 10% more and the other half reported 5% to 10% gains. In terms of predictions,

predictions where the it was confident enough to assign a single label.

Distinguishing next between chemistry originating from external and internal sources ([Table 1](#_bookmark12)), it can be observed that both showed a similar picture, being positive and in a similar range, and neither of the two being consistently higher ([Table 2](#_bookmark15)). The mean values were generally higher than the median values.

Next, the variation of the ΔCE values for a range of datasets was

investigated. The datasets originated from sources having varied com-

positions with the goal of effectively capturing wide regions of small- molecule drug-like space to obtain a general view. [Fig. 6](#_bookmark18) details the

partner individually, through displaying the distribution of ΔCE over behavior of the individual datasets, both unlabeled and labeled for each

the tasks. A decrease in uncertainty in the multipartner setting appears as a distribution above the zero line, and vice versa, any increase in

[Table 2](#_bookmark15), the median ΔCE values were predominantly positive, indicating uncertainty is reflected in distributions below the zero line. In line with

a decrease in uncertainty in the multipartner setting. Some observations can be made. First, there were markedly higher gains in unlabeled space

resentative ΔCE values, such as ChEMBL 25 - MELLODDY-TUNER, with compared to labeled space. Comparing an unlabeled dataset with rep-

the test set, nine out of ten partners (except partner H) reported quartiles 1 to 3 to all be higher in the unlabeled space. The gains in the unlabeled space were also fairly consistent between the different datasets, even between internal and external chemistry, as compounds in the training, validation and test folds were more in line with other unlabeled datasets

than with the test set. In particular, whereas the labeled space ΔCE val-

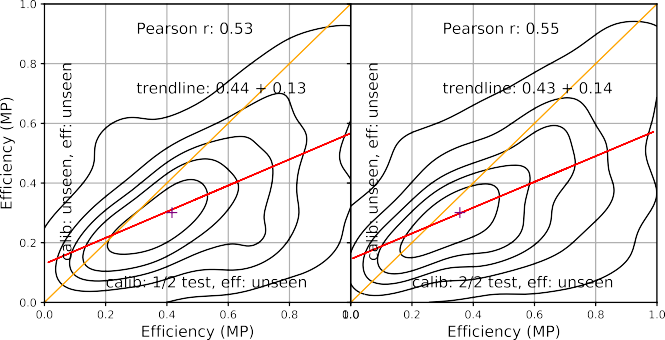
ues had interquartile ranges containing zero and medians around zero,

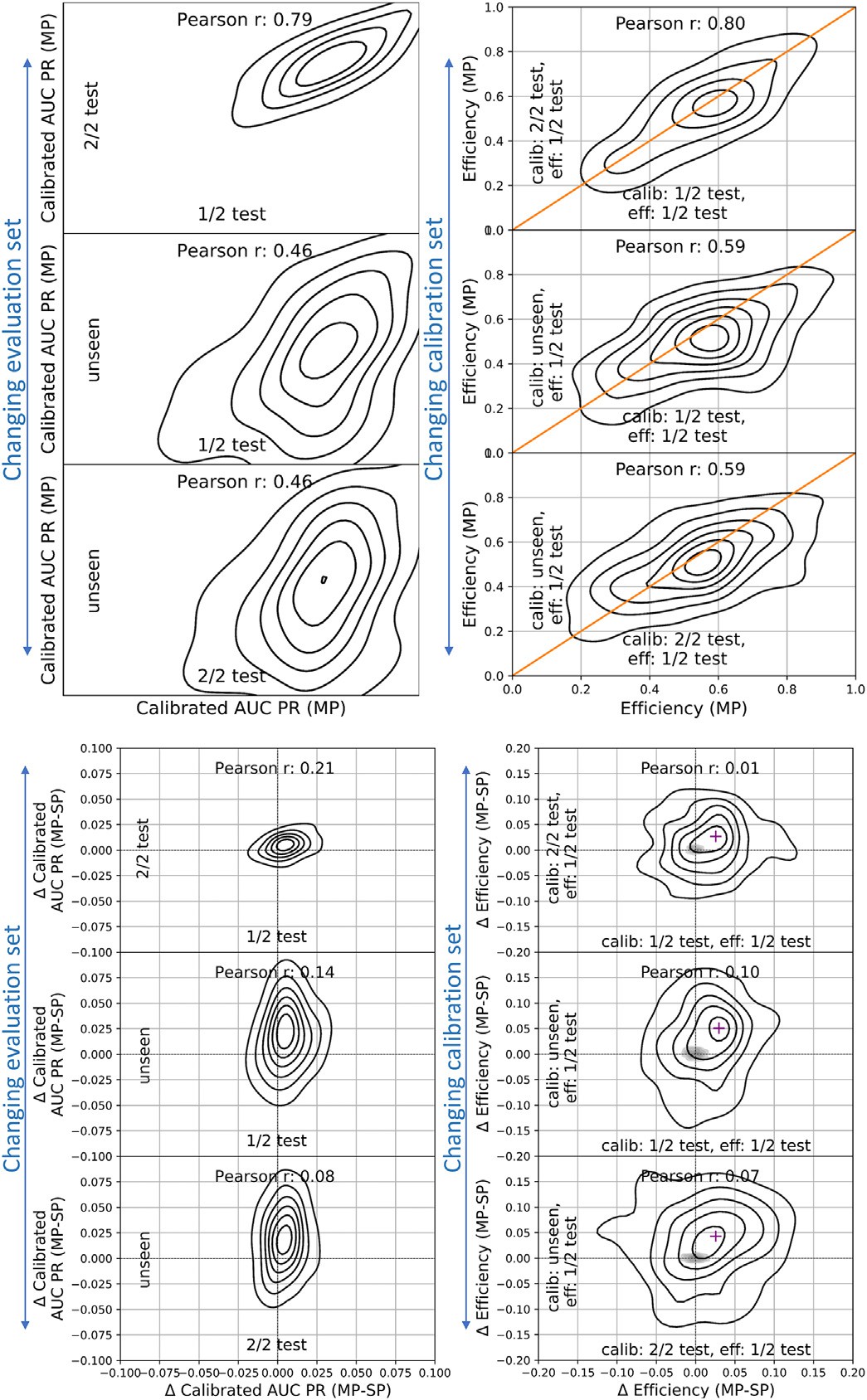
this was not the case for the datasets in unlabeled space, which more often tended to have positive medians and interquartile ranges with the first quartile close to zero. For the data that the models had seen dur- ing training (training folds 1–3 and in case of Phase 2 models also the validation fold), interquartile ranges containing zero and median val- ues around zero might be expected, but this was also the case for the test set, despite it being unseen during training. Although the test set

**Table 2**

Overview of the delta conformal eﬃciency (ΔCE) for all partners A-J, for unlabeled datasets (See [Table 1](#_bookmark12) for details).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Partner | Median ΔCE |  |  |  | Mean ΔCE |  |  |
|  | External chemistry | Internal chemistry | Any chemistry |  | External chemistry | Internal chemistry | Any chemistry |
| A | 9.3% | 7.1% | 8.7% |  | 12.4% | 9.8% | 11.7% |
| B | 8.0% | 7.4% | 7.8% |  | 10.2% | 8.3% | 9.7% |
| C | 4.3% | 4.9% | 4.5% |  | 6.0% | 6.2% | 6.1% |
| D | 6.5% | 5.6% | 6.2% |  | 7.9% | 6.9% | 7.6% |
| E | 4.4% | 4.3% | 4.4% |  | 7.7% | 4.9% | 6.9% |
| F | 10.5% | 8.1% | 9.7% |  | 11.6% | 5.4% | 9.9% |
| G | 4.8% | 5.8% | 4.9% |  | 5.4% | 4.1% | 5.1% |
| H | 0.4% | 1.4% | 0.7% |  | 1.1% | 2.5% | 1.5% |
| I | 2.2% | 3.8% | 2.7% |  | 3.1% | 4.4% | 3.4% |
| J | 5.4% | 6.4% | 5.7% |  | 7.0% | 6.9% | 6.9% |
| Mean | 5.6% | 5.5% | **5.5%** |  | 7.2% | 5.9% | 6.9% |



**Fig. 5.** Comparison of conformal eﬃciency calculation when calibrating on part of the test fold and calculating eﬃciency on the ‘unseen’ dataset, to the situa- tion using the ‘unseen’ dataset for both. Plots show the density of tasks (kernel

density estimation). The purple plus sign (+) marks the maximal density. The

linear trendline is shown in red. The diagonal is shown in orange.

**Fig. 4.** Correlation of conformal eﬃciency with performance metrics requiring labels. Left: linear correlation between the predictive performance (calibrated AUC PR) on different evaluation sets. Right: linear correlation between the con- formal eﬃciency with different calibration sets. Top: metrics. Bottom: metric deltas. Plots show the density of tasks (kernel density estimation). The gray- shaded contour plot in the bottom right shows the distribution of maximal den- sities from the permutation experiments where the set of single-partner and mul- tipartner model predictions for a task were interchanged randomly. The purple

plus sign (+) marks the maximal density.

other four internal labeled datasets, its ΔCE values were still more in line showed a somewhat greater decrease of uncertainty compared to those

with those labeled datasets than with datasets where no label is avail- able. This observation indicates that the test set was not representative of unseen types of chemistry.

Although the gains in the unlabeled space were fairly consistent be- tween datasets, some differences could be noticed to recur for most part- ners. For instance, the DrugSpaceX dataset tended to show lower values, with nine out of ten partners reporting the lowest gains across all un- labeled datasets. On the other hand, the FDA-approved small-molecule drugs were among the datasets showing the largest decrease in uncer-

tainty, with seven out of ten partners reporting the highest ΔCE val-

dataset can affect the observed ΔCE. Indeed, a relationship between the ues ([Fig. 6](#_bookmark18)). Such findings indicate that some property of the unlabeled size of the molecules and the ΔCE was found. Unlabeled datasets with

tended to show higher ΔCE values than datasets with larger molecules, smaller molecules, such as the FDA-approved small-molecule drugs,

such as DrugSpaceX. [Fig. 7](#_bookmark19) shows that this relationship held for interme- diate sizes as well. Similar observations can be made by considering the

ΔPlatt-scaled entropy, such as the consistent decrease of uncertainty for

unlabeled space, a markedly different behavior between unlabeled and

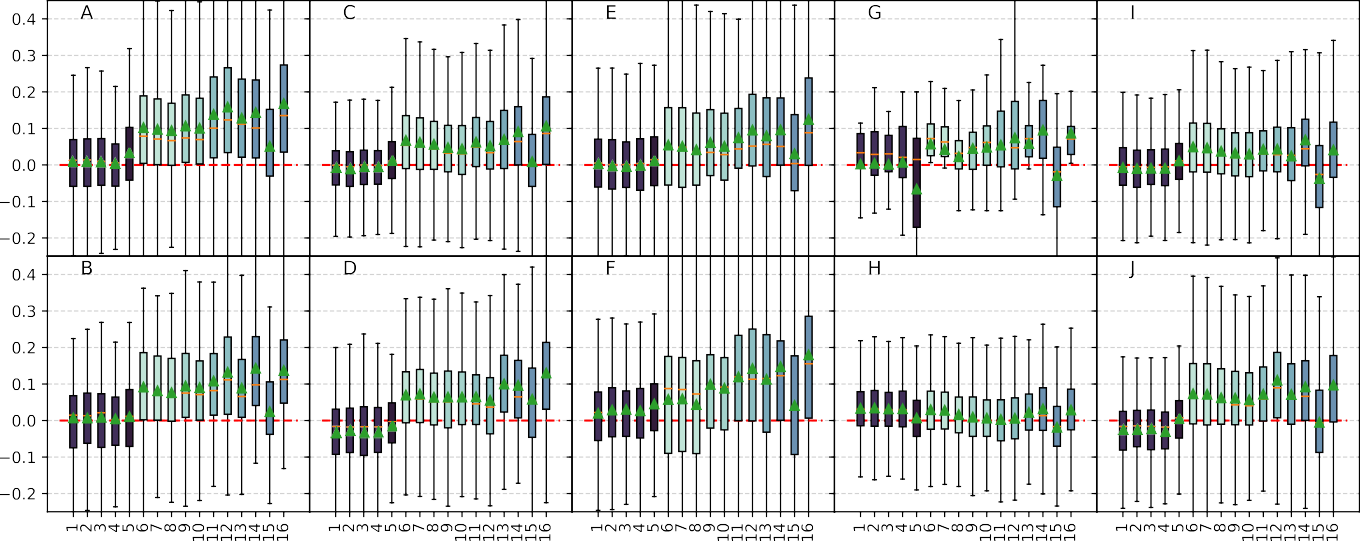
labeled space and the finer distinctions between the unlabeled datasets (Fig. S6).

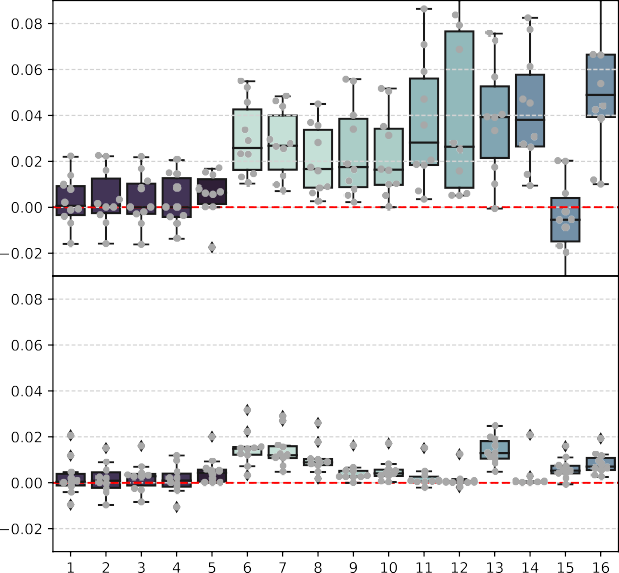
The overall ΔCE can be split into two parts: one related to the inac-

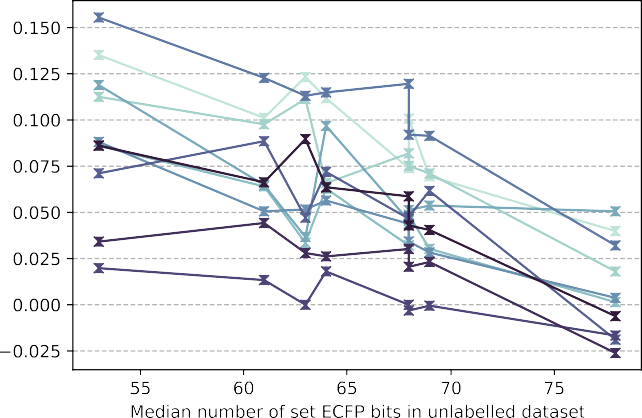
tive predictions, and one related to the active predictions ([Fig. 2](#_bookmark13)). [Fig. 8](#_bookmark19)

not exclusively to the ΔCE. This was somewhat expected considering shows that the inactive predictions contributed more, but importantly

the general class balance of the tasks, which tended to be dominated by inactives. Hence, both actives and inactives will be predicted more confidently in the multipartner setting.

**Fig. 6.** Predictive uncertainty analysis *via* the distribution of delta (multipartner minus single-partner) conformal eﬃciency (ΔCE) for several datasets over all tasks for all partners A-J. Lighter shades represent the unlabeled datasets, darker shades the labeled datasets ([Table 1](#_bookmark12)).





terms of number of set ECFP bits), and the delta conformal eﬃciency (ΔCE). **Fig. 7.** Relationship between the size of the molecules in unlabeled datasets (in

The lines indicate the various partners.

The possible sensitivity of uncertainty metrics on the hyperparam- eters may invoke questions whether the comparison of these metrics between the single- and multipartner models trained with different hy- perparameters can be meaningful. To ensure that any observed trends were not spuriously due to hyperparameters, single- and multipartner models from several hyperparameter combinations were investigated. Similar trends were observed when (i) varying the single-partner hyper-

a limited degree (*<*0.5% AUC PR overall); (ii) switching from Phase 2 parameters in local training while deteriorating model performance by

to Phase 1 models. Figs. S7 and S8 provide detailed results.

# Discussion

The main research question of this work is whether federated mul- titask learning extends the applicability domain of property prediction models compared to single-partner settings. Based on intermediate fed- erated models from the MELLODDY project, this work proposes the con- formal eﬃciency as a means to measure the applicability domain exten- sion through predictive uncertainty differences.

Considering that the applicability domain can be defined as the chemical space in which the model makes predictions with suﬃcient

**Fig. 8.** Median delta (multi- minus single-partner) conformal eﬃciency (ΔCE) for several datasets over all tasks for all pharma partners. Top: ΔCE consid- ering only inactive predictions. Bottom: ΔCE considering only active predic-

tions. Lighter shades represent the unlabeled datasets, darker shades the labeled datasets ([Table 1](#_bookmark12)).

reliability [[28](#_bookmark32),[29](#_bookmark34)], an extension of the applicability domain can be re- lated to a decrease in predictive uncertainty. Several studies have shown the good performance of uncertainty measures for applicability domain estimation [[26](#_bookmark29),[66–72](#_bookmark57)], on par or superior to other common approaches as distance-based or geometric techniques in the input space [[23]](#_bookmark25) or latent space of the neural network [[73]](#_bookmark66). The former would require ele- vated access to the set of training data across partners, prompting pri- vacy concerns. Conceptually, the applicability domain pertains to the epistemic uncertainty as samples inside the applicability domain can still be predicted with high uncertainty if they are inherently noisy or ambiguous.

A number of authors have related the applicability domain with con- formal predictors in binary classification [[26](#_bookmark29),[74](#_bookmark67),[75](#_bookmark69)]. More specifically, one interpretation is to consider prediction sets with both class labels to

be inherently noisy, hence of high aleatoric uncertainty, but inside the applicability domain, while empty prediction sets can be considered of high epistemic uncertainty and outside of the applicability domain since they are different from the known instances of the existing classes [[74]](#_bookmark67). In the current study, contrastingly, a change in uncertainty measured through the eﬃciency was linked directly to the applicability domain extension, without making the distinction whether a high uncertainty prediction was associated with none or both class labels. Such appli- cability domain obtained through conformal predictors and its depen- dency on a user-specified error level differs conceptually from the appli- cability domain obtained with distance-based or geometric techniques in the input space.

In a sense, the objective of estimating the applicability domain ex- tension based on uncertainty metrics is somewhat related to other ap- proaches to estimate the model quality and generalization without re- quiring access to labels. The Predicting Generalization in Deep Learn- ing (PGDL) competition is an example where complexity measures, data augmentation techniques, and representation analyses are studied that predict the level of generalization for a model, likewise without requir- ing labels [[76]](#_bookmark70). Another example is based on the decomposition of model weight matrices [[77]](#_bookmark71).

To select an uncertainty metric, in §3.2. conformal eﬃciency was compared to other established uncertainty metrics such as predictive entropy, variance, and IG. Unlike conformal eﬃciency and entropy that behaved intuitively, variance and IG showed counterintuitive behavior and were therefore deprioritized as a suitable uncertainty metric. The choice for CE as the final metric was further favored by its flexibility as distribution-free method and its use in industrial settings [[30–32,63](#_bookmark36)]. Nevertheless, the main finding of an extended AD for the federated model was also confirmed with the Platt-scaled entropy.

A decrease in predictive uncertainty leads to more accurate under- lying predictions with well-calibrated models, but in other situations models may display an unjustified or ‘empty’ decrease of uncertainty that does not correspond to predictive gains. To address this challenge, in §3.3. a relationship was established between conformal eﬃciency and performance metrics requiring labels. Theoretically, a strong link with the TPR and TNR was demonstrated that is valid at low significance lev- els, providing novel insight into the nature of the eﬃciency. Empirically, the relationship between conformal eﬃciency and calibrated AUC PR at task level was showcased for several settings in which the exchange-

ability assumption is increasingly challenged. A decent correlation (r *>*

0.50) was found even in the most challenging of settings, with an ‘un-

seen’ dataset that originated after the model training and did not con- tain any scaffolds included in the training, validation, or test sets. This ‘unseen’ dataset was arguably the best possible approximation of unla- beled space within the boundaries of the available, realistic, industrial data. Thus, the uncertainty from CE seemed justified as CE continued to correlate with established predictive performance metrics, even where exchangeability could no longer be assumed.

Exchangeability is typically confirmed by evaluating the validity on a held-out set. One of the fundamental challenges faced in this work, is that potential benefits of federated learning may well manifest in regions of chemical space, where no labels are available to the eval- uating partner because they are only explored by other partners. This space was represented by composing so-called unlabeled datasets based on varied drug-like libraries from different sources. Such absence of labels precludes any calculation of the validity. Conformal prediction formally assumes the exchangeability of the calibration and test sets, in which case it guarantees to adhere to a user-specified error level. In this work, conformal prediction was intentionally applied on com- pounds lying outside of the internally labeled chemical space of every participating pharma partner and thus were expected to be different from the calibration set. No formal guarantee of exchangeability, nor of error levels on the unlabeled datasets can therefore be claimed. The less plausible the exchangeability assumption, the more loose becomes the interpretation of conformal eﬃciency as a metric of AD. Importantly,

this loss of correlation is not abrupt, but gradual, reflecting a gradual erosion of the plausibility of exchangeability and interpretability of eﬃ- ciency. Even at maximally challenging conditions, as closely as possible resembling unlabeled space, the correlation with performance metrics requiring labels, and hence its interpretability, had eroded but was not lost.

Undoubtedly there is a tension between the formal guarantees on error rates that accompany a perfect exchangeability and the erosion of the exchangeability assumption specifically for unlabeled compounds that is relevant for practical drug discovery settings. One of such prac- tical settings could be virtual screening to discover novel chemical scaf- folds with activity. Federated models hold promise here, considering the chemical space covered by a federated model is the combination of chemical spaces of the individual, contributing partners, and therefore indications of activity might appear in unexplored regions of chemical space due to other partners’ data. Due to privacy constraints, each party only has access to the labels of its own compound libraries. This con- strains the approach to uncertainty metrics that do not require labels derived from meticulous experimentation.

of ΔCE was investigated. Varying calibration and evaluation datasets After having established the CE as uncertainty metric, the behavior for ΔCE yielded only very low correlations. Importantly, the task- aggregated ΔCE maintained its direction, prompting the use of the task- aggregated ΔCE to investigate the extension of the AD for the federated

of the ΔCE was found to be similar to that of the Δcalibrated AUC PR models. It should be noted that this sensitivity to the dataset at hand and therefore should not be seen as a comparative disadvantage of ΔCE.

Additionally, uncertainty metrics such as CE explicitly take information of the predicted compounds into account, in contrast to labeled metrics where the best estimation of model predictive performance on held-out data is the value resulting from evaluating the test set, regardless of the nature of the predicted compounds.

In §3.4. the main research question of the extension of the AD for the

uncertainty for the federated models, the positive task-aggregated ΔCE federated models was evaluated. While a portion of the tasks increased

indeed indicated an overall decrease in uncertainty compared to the single-partner model, hence suggesting extension of the applicability do- main as beneficial effect. Such decrease in uncertainty was particularly present where no labels were available, regardless of the source of chem-

ten MELLODDY partners. Based on the observation that the ΔCE might istry, including internal chemistry. The effect was consistent across the

overestimate the performance gain by 25% (see above for discussion on [Fig. 5](#_bookmark17)), a 4.1% increase in conformal eﬃciency was estimated to be more reflective of the predictive performance gain than the 5.5% me-

The findings were confirmed with the ΔPlatt-scaled entropy similarly dian conformal eﬃciency increase reported across partners ([Table 2](#_bookmark15)).

showing a task-aggregated decrease in uncertainty.

One of the other aspects that became apparent is that the ΔCE val-

ues of the test set were found to be generally closer to those of the

training set and validation set than to the datasets where no labels were available. This can serve as an additional indication of the erosion of the exchangeability assumption for unlabeled compounds. The follow- ing considerations may help explain this observation.

First, the test set can be similar to the training and validation set due to limitations related to splitting an internal dataset into five folds. The underlying idea of the scaffold-based splitting approach is to separate distinct areas of chemical space and to assign them to different folds thus avoiding overconfident predictive performance estimates. While scaffold-network-based splitting achieves far better separation of similar molecules into different folds than random splitting, some structurally very similar compounds might still be assigned to different folds due to not sharing the same scaffold [[39]](#_bookmark51). Switching to other splitting schemes such as sphere exclusion clustering was not expected to fundamentally alter the picture based on its limited performance gain over scaffold- network-based splitting [[39]](#_bookmark51). While scaffold-based splitting challenges

the exchangeability between the folds, the relative nearness of ΔCE be-

tween the test fold and the other labeled datasets compared to unlabeled datasets indicated that the effect is relatively small compared to the ero- sion of the exchangeability assumption for unlabeled datasets.

Secondly, the internal data at a pharmaceutical company might not be representative of the drug-like chemical space at large. The devia- tion between the test set (low uncertainty decrease) and the unlabeled compounds in the test fold (high uncertainty decrease) suggests that the occurrence of labels is biased to certain regions that are already informed by the single-partner model. Hence, the probability of a com- pound being tested in an assay is not random. Reasons include known or suspected activity against a similar target or membership of a commonly used screening library, and series bias or analogue bias due to drug dis- covery focusing on a relatively narrow set of related compounds during chemical synthesis and subsequent testing [[78]](#_bookmark72). Additionally, factors re- lated to the composition of the internal compound library might impede the creation of a fold that represents truly unseen chemistry. Since inter- nal libraries only cover a fraction of the total drug-like chemical space, and typically contain many related compounds through the occurrence of chemical series, they cannot be expected to be general. All these fac- tors and others [[79]](#_bookmark73) may contribute to the observed difference between labeled and unlabeled datasets.

As discussed, the predictive uncertainty from different unlabeled datasets is fairly consistent and exceeds that of the labeled datasets. The observation that both external and internal chemistry decrease in uncertainty, suggests, perhaps contrary to the expectation [[80]](#_bookmark75), that it is possible to acquire knowledge about one’s own compounds on one’s

ner learning. The observed stark difference in ΔCE (and ΔPlatt-scaled own assays through the other partners’ data in the federated multipart-

entropy) for the labeled and unlabeled datasets supports the choice to explicitly investigate unlabeled data separately, where validity calcula- tion is out of reach.

Despite this general picture, differences between the unlabeled datasets existed. The ‘FDA-approved small-molecule drugs’ dataset gen- erally showed a larger uncertainty decrease in the multipartner setup than the other unlabeled datasets, while the ‘DrugSpaceX’ dataset showed relatively lower uncertainty decreases. One might consider the explanation that such FDA-approved small-molecule drugs likely exist in the internal libraries of many partners as reference compounds, and as such, have many activity measurements associated with them, con- tributing to their ability to exchange information on different tasks dur- ing the federated multipartner learning. Since the DrugSpaceX dataset consists of virtual compounds with virtually zero overlap with pharma libraries, one might attribute the lower uncertainty decrease to the lack of overlap. However, in opposition to this explanation, the other virtual

tionship between the ΔCE and the data sets’ molecular size in terms of library (SCUBIDOO) did not exhibit such behavior, and a strong rela- median number of set ECFP bits was found. When switching from ΔCE to ΔPlatt-scaled entropy, similar observations such as the markedly differ-

ent behavior of unlabeled compared to labeled space and the finer dis- tinctions between the unlabeled datasets were made. While this show- cased that shifted predictive probabilities play an essential part, the

ΔPlatt-scaled entropy still required a calibration set hereby triggering

exchangeability considerations similar to those for ΔCE.

# Conclusions

Federated learning holds the promise of substantial predictive per- formance gain in the chemical space informed by other partners’ data, where no labels are available to the evaluating partner. Any evaluation requiring labels, would trap the evaluating partner in its own space, blind for any changes beyond that space. Since uncertainty metrics do not require labels, conformal eﬃciency was proposed as a metric to estimate a model’s applicability domain, after establishing a theoreti- cal and empirical relationship with established predictive performance metrics requiring labels. Such relationship was shown for low signifi- cance levels, and for situations where the exchangeability assumption is

maximally challenged. This suggested that decreases in uncertainty as measured with conformal eﬃciency correspond to predictive benefits, even for unlabeled datasets.

An overall increase of the ΔCE was observed in the unlabeled space.

Such increase remarkably occurred for both internal and external chem-

istry and could not be recovered from the labeled test fold, indicating that this fold was not representative of unlabeled types of chemistry.

The decrease in uncertainty suggested an extension of the applicabil- ity domain of the federated model and qualified as a tangible benefit of federated learning. On average, MELLODDY partners reported a median

increase in ΔCE of the federated over the single-partner model of 5.5%

(with increases up to 9.7%), signifying that the models stemming from

the federated learning will produce up to 10% more low uncertainty (i.e., single class label) predictions.

As an outlook, the insight into the conformal eﬃciency at these low significance levels and its establishment as applicability domain metric could extend its usage in the future, particularly for large scale multitask model comparisons.

In conclusion, based on interim MELLODDY results, the first indi- cation is presented that privacy-preserving federated machine learning across massive drug-discovery datasets from ten pharma partners indeed extends the applicability domain of property prediction models.

# Briefs

Based on interim results from the MELLODDY project, the first indi- cation is presented that privacy-preserving federated machine learning across massive drug-discovery datasets from ten pharma partners indeed extends the applicability domain of property prediction models.

# Data and software availability

The massive underlying datasets are proprietary and cannot be shared, but code for the analysis, particularly relating to the calibration [and calculation of uncertainty metrics can be found at https://github. com/melloddy/conformal\_eﬃciency. A small-scale public dataset on](https://github.com/melloddy/conformal_efficiency) carboxylic acids and other acids to reproduce selected results from the SI has been provided at <https://zenodo.org/communities/melloddy/>.

# Funding sources

This project has received funding from the Innovative Medicines Ini- tiative 2 Joint Undertaking under grant agreement N° 831472. This Joint Undertaking receives support from the European Union’s Hori- zon 2020 research and innovation program and EFPIA. This research received funding from the Flemish Government (AI Research Program). Y.M., A.A., J.S., M.O., and R.L. are aﬃliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

# Declaration of Competing Interest

The authors declare that they have no known competing financial in- terests or personal relationships that could have appeared to influence the work reported in this paper. W.H., H.C., A.P., N.S., A.S., L.H., A.Af., L.M., A.Z., L.F. were employees of Janssen, Bayer, Novartis, Boehringer Ingelheim, Astellas, AstraZeneca, Amgen and Merck, respectively, dur- ing the study.

# CRediT authorship contribution statement

**Wouter Heyndrickx:** Conceptualization, Methodology, Investiga- tion, Data curation, Formal analysis, Software, Writing – original draft, Writing – review & editing. **Adam Arany:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – review & editing. **Jaak**

**Simm:** Conceptualization, Methodology, Formal analysis, Writing – re- view & editing. **Anastasia Pentina:** Data curation, Investigation, Writ- ing – original draft, Writing – review & editing. **Noé Sturm:** Data cu- ration, Investigation, Writing – review & editing. **Lina Humbeck:** Data curation, Investigation, Writing – review & editing. **Lewis Mervin:** Data curation, Writing – review & editing. **Adam Zalewski:** Data curation, Investigation, Writing – review & editing. **Martijn Oldenhof:** Concep- tualization, Methodology, Formal analysis, Writing – review & editing. **Peter Schmidtke:** Data curation, Investigation, Writing – review & edit- ing. **Lukas Friedrich:** Data curation, Investigation, Writing – review & editing. **Regis Loeb:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Arina Afanasyeva:** Investigation, Data cu- ration, Writing – review & editing. **Ansgar Schuffenhauer:** Conceptu- alization, Methodology, Writing – review & editing. **Yves Moreau:** Con- ceptualization, Methodology, Formal analysis, Writing – review & edit- ing. **Hugo Ceulemans:** Conceptualization, Methodology, Formal anal- ysis, Funding acquisition, Writing – original draft, Writing – review & editing.

# Data availability

The code for the analysis (https://github.com/melloddy/conformal\_ eﬃciency) and a public dataset to reproduce results from the SI (https://zenodo.org/communities/melloddy/) has been made available.

# Acknowledgment

We are grateful to Luc Geeraert for excellent scientific writing assis- tance, and to Thanh Le Van for stimulating discussions.

# Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.ailsci.2023.100070](https://doi.org/10.1016/j.ailsci.2023.100070).

# References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S. Applications of machine learn- ing in drug discovery and development. Nat Rev Drug Discov 2019;18:463–77. doi:[10.1038/s41573-019-0024-5](https://doi.org/10.1038/s41573-019-0024-5).
2. Engels MFM, Gibbs AC, Jaeger EP, Verbinnen D, Lobanov VS, Agrafiotis DK. A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. J Chem Inf Model 2006;46:2651–60. doi:[10.1021/ci600219n](https://doi.org/10.1021/ci600219n).
3. Kogej T, Blomberg N, Greasley PJ, Mundt S, Vainio MJ, Schamberger J, Schmidt G, Hüser J. Big pharma screening collections: more of the same or unique libraries? the AstraZeneca-Bayer Pharma AG case. Drug Discov Today 2013;18:1014–24. doi:[10.1016/j.drudis.2012.10.011](https://doi.org/10.1016/j.drudis.2012.10.011).
4. Schamberger J, Grimm M, Steinmeyer A, Hillisch A. Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering. Drug Dis- cov Today 2011;16:636–41. doi:[10.1016/j.drudis.2011.04.005](https://doi.org/10.1016/j.drudis.2011.04.005).
5. Bosc N, Felix E, Arcila R, Mendez D, Saunders MR, Green DVS, Ochoada J, Shelat AA, Martin EJ, Iyer P, Engkvist O, Verras A, Duffy J, Burrows J, Gardner JMF, Leach AR. MAIP: a web service for predicting blood-stage malaria inhibitors. J Cheminform 2021;13:13. doi:[10.1186/s13321-021-00487-2](https://doi.org/10.1186/s13321-021-00487-2).
6. Verras A, Waller CL, Gedeck P, Green DVS, Kogej T, Raichurkar A, Panda M, Shelat AA, Clark J, Guy RK, Papadatos G, Burrows J. Shared consensus machine learning models for predicting blood stage malaria inhibition. J Chem Inf Model 2017;57:445–53. doi:[10.1021/acs.jcim.6b00572](https://doi.org/10.1021/acs.jcim.6b00572).
7. [Brendan McMahan H, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communi- cation-eﬃcient learning of deep networks from decentralized data. In: Proceedings of the 20th international conference on artificial intelligence and statistics, AISTATS 2017, 54; 2017. p. 1273–82.](http://refhub.elsevier.com/S2667-3185(23)00014-4/sbref0007)
8. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmen- tation. Brainlesion 2019;11383:92–104. doi:[10.1007/978-3-030-11723-8\_9](https://doi.org/10.1007/978-3-030-11723-8_9).
9. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applica- tions. ACM Trans Intell Syst Technol 2019;10:1–19. doi:[10.1145/3298981](https://doi.org/10.1145/3298981).
10. Ruder S. An overview of multi-task learning in deep neural networks, ArXiv. (2017) arXiv: 1706.05098. <http://arxiv.org/abs/1706.05098>.
11. [Caruana R. Multi-task learning. Mach Learn 2018;28:41–75. doi:10.1023/ 1007379606734.](https://doi.org/10.1023/\penalty%20-\%40M%201007379606734)
12. [Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner J, Ceulemans H, Hochre- iter S. Deep learning as an opportunity in virtual screening. Adv Neural Inf Process Syst 2014;27:1–9.](http://refhub.elsevier.com/S2667-3185(23)00014-4/sbref0012)
13. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. Mas- sively multitask networks for drug discovery, ArXiv. (2015) arxiv ID: 1502.02072. <http://arxiv.org/abs/1502.02072>.
14. Kearnes S, Goldman B, Pande V. Modeling industrial ADMET data with multitask networks, ArXiv. (2016) arXiv ID: 1606.08793. <http://arxiv.org/abs/1606.08793>.
15. Lenselink EB, Ten Dijke N, Bongers B, Papadatos G, Van Vlijmen HWT, Kowalczyk W, Ijzerman AP, Van Westen GJP. Beyond the hype: deep neural networks outper- form established methods using a ChEMBL bioactivity benchmark set. J Cheminform 2017;9:1–14. doi:[10.1186/s13321-017-0232-0](https://doi.org/10.1186/s13321-017-0232-0).
16. Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V. Demystifying multitask deep neu- ral networks for quantitative structure-activity relationships. J Chem Inf Model 2017;57:2490–504. doi:[10.1021/acs.jcim.7b00087](https://doi.org/10.1021/acs.jcim.7b00087).
17. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, Clev- ert DA, Hochreiter S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci 2018;9:5441–51. doi:[10.1039/c8sc00148k](https://doi.org/10.1039/c8sc00148k).
18. Wenzel J, Matter H, Schmidt F. Predictive multitask deep neural network mod- els for ADME-Tox properties: learning from large data sets. J Chem Inf Model 2019;59:1253–68. doi:[10.1021/acs.jcim.8b00785](https://doi.org/10.1021/acs.jcim.8b00785).
19. Sturm N, Mayr A, Van TLe, Chupakhin V, Ceulemans H, Wegner J, Golib-Dzib JF, Jeliazkova N, Vandriessche Y, Böhm S, Cima V, Martinovic J, Greene N, Vander Aa T, Ashby TJ, Hochreiter S, Engkvist O, Klambauer G, Chen H. Industry-scale application and evaluation of deep learning for drug target prediction. J Cheminform 2020;12:1–13. doi:[10.1186/s13321-020-00428-5](https://doi.org/10.1186/s13321-020-00428-5).
20. Heyndrickx W, Mervin L, Morawietz T, Sturm N, Friedrich L, Zalewski A, Pentina A, Humbeck L, Oldenhof M, Niwayama R, Schmidtke P, Simm J, Arany A, Drizard N, Jabal R, Afanasyeva A, Loeb R, Harnqvist S, Holmes M, Pejo B, Telenczuk M, Holway N, Rieke N, Zumsande F, Clevert D, Krug M, Green D, Ertl P, An- tal P, Marcus D, Do Huu N, Fuji H, Pickett S, Acs G, Boniface E, Beck B, Sun Y, Gohier A, Engkvist O, Göller A.H, Moreau Y, Galtier M.N, Ceulemans H. MEL- LODDY : cross pharma federated learning at unprecedented scale unlocks benefits [in QSAR without compromising proprietary information, 2022. https://chemrxiv. org/engage/chemrxiv/article-details/6345c0f91f323d61d7567624.](https://chemrxiv.org/engage/chemrxiv/article-details/6345c0f91f323d61d7567624)
21. Oldenhof M, Ács G, Pejó B, Schuffenhauer A, Holway N, Sturm N, Dieckmann A, Fortmeier O, Boniface E, Mayer C, Gohier A, Schmidtke P, Niwayama R, Kopecky D, Mervin L, Rathi PC, Friedrich L, Formanek A, Antal P, Rahaman J, Zalewski A, Heyn- drickx W, Oluoch E, Stößel M, Vančo M, Endico D, Gelus F, de Boisfossé T, Darbier A, Nicollet A, Blottière M, Telenczuk M, Nguyen VT, Martinez T, Boillet C, Moutet K, Picosson A, Gasser A, Djafar I, Arany Á, Simm J, Moreau Y, Engkvist O, Ceule- mans H, Marini C, Galtier M. Industry-scale orchestrated federated learning for drug discovery, ArXiv. (2022) arXiv ID: 2210.08871. <http://arxiv.org/abs/2210.08871>.
22. Breiman L. Random forests. Mach Learn 2001;45:5–32. doi:10.1023/A: 1010933404324.
23. Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncer- tainty quantification in drug design. Drug Discov Today 2021;26:474–89. doi:[10.1016/j.drudis.2020.11.027](https://doi.org/10.1016/j.drudis.2020.11.027).
24. Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world, 2005. doi:[10.1007/b106715](https://doi.org/10.1007/b106715).
25. Cortés-Ciriano I, Bender A. Concepts and applications of conformal predic- tion in computational drug discovery, ArXiv. (2019) arXiv ID: 1908.03569. <http://arxiv.org/abs/1908.03569>.
26. Norinder U, Carlsson L, Boyer S, Eklund M. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicabil- [ity domain determination. J Chem Inf Model 2014;54:1596–603. doi:10.1021/ ci5001168.](https://doi.org/10.1021/\penalty%20-\%40M%20ci5001168)
27. Norinder U, Spjuth O, Svensson F. Synergy conformal prediction applied to large

- scale bioactivity datasets and in federated learning. J Cheminform 2021:1–11. doi:[10.1186/s13321-021-00555-7](https://doi.org/10.1186/s13321-021-00555-7).

1. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Ja- worska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, Van De Sandt JJM, Tong W, Veith G, Yang C. Current status of methods for defining the applicabil- ity domain of (quantitative) structure-activity relationships. ATLA Altern Lab Anim 2005;33:155–73. doi:[10.1177/026119290503300209](https://doi.org/10.1177/026119290503300209).
2. Hanser T, Barber C, Marchaland JF, Werner S. Applicability domain: to- wards a more formal definition. SAR QSAR Environ Res 2016;27:893–909. doi:[10.1080/1062936X.2016.1250229](https://doi.org/10.1080/1062936X.2016.1250229).
3. Sun J, Carlsson L, Ahlberg E, Norinder U, Engkvist O, Chen H. Apply- ing Mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. J Chem Inf Model 2017;57:1591–8. doi:[10.1021/acs.jcim.7b00159](https://doi.org/10.1021/acs.jcim.7b00159).
4. Morger A, Mathea M, Achenbach JH, Wolf A, Buesen R, Schleifer KJ, Landsiedel R, Volkamer A. KnowTox: pipeline and case study for confident prediction of po- tential toxic effects of compounds in early phases of development. J Cheminform 2020;12:1–17. doi:[10.1186/s13321-020-00422-x](https://doi.org/10.1186/s13321-020-00422-x).
5. Garcia de Lomana M, Morger A, Norinder U, Buesen R, Landsiedel R, Volkamer A, Kirchmair J, Mathea M. ChemBioSim: enhancing conformal prediction of *in vivo* toxicity by use of predicted bioactivities. J Chem Inf Model 2021;61:3255–72. doi:[10.1021/acs.jcim.1c00451](https://doi.org/10.1021/acs.jcim.1c00451).
6. Morger A, Garcia de Lomana M, Norinder U, Svensson F, Kirchmair J, Mathea M, Volkamer A. Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data. Sci Rep 2022;12:1–13. doi:[10.1038/s41598-022-09309-3](https://doi.org/10.1038/s41598-022-09309-3).
7. MELLODDY-TUNER, (2021). <https://github.com/melloddy/MELLODDY-TUNER>.
8. Landrum G. RDKit: open-source cheminformatics software, (2021). <http://www.rdkit.org/>.
9. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54. doi:[10.1021/ci100050t](https://doi.org/10.1021/ci100050t).
10. Varin T, Gubler H, Parker CN, Zhang JH, Raman P, Ertl P, Schuffenhauer A. Com- pound set enrichment: a novel approach to analysis of primary HTS data. J Chem Inf Model 2010;50:2067–78. doi:[10.1021/ci100203e](https://doi.org/10.1021/ci100203e).
11. Kruger F, Stiefl N, Landrum GA. RdScaffoldNetwork: the Scaffold net- work implementation in RDKit. J Chem Inf Model 2020;60:3331–5. doi:[10.1021/acs.jcim.0c00296](https://doi.org/10.1021/acs.jcim.0c00296).
12. Simm J, Humbeck L, Zalewski A, Sturm N, Heyndrickx W, Moreau Y, Beck B, Schuf- fenhauer A. Splitting chemical structure data sets for federated privacy-preserving machine learning. J Cheminform 2021;13:1–14. doi:[10.1186/s13321-021-00576-2](https://doi.org/10.1186/s13321-021-00576-2).
13. Humbeck L, Morawietz T, Sturm N, Zalewski A, Harnqvist S, Heyndrickx W, Holmes M, Beck B. Don’t overweight weights: evaluation of weighting strate- gies for multi-task bioactivity classification models. Molecules 2021;26:6959. doi:[10.3390/molecules26226959](https://doi.org/10.3390/molecules26226959).
14. Arany A, Simm J, Oldenhof M, Moreau Y. SparseChem: fast and accurate ma- chine learning model for small molecules, ArXiv. (2022) arXiv ID: 2203.04676. <http://arxiv.org/abs/2203.04676>.
15. [Platt J. Probabilistic outputs for support vector machines and comparisons. Advances in Large Margin. Classifiers 1999:61–74.](http://refhub.elsevier.com/S2667-3185(23)00014-4/sbref0042)
16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011:6. <https://arxiv.org/abs/1201.0490>.
17. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform 2019;11:1–16. doi:[10.1186/s13321-018-0325-4](https://doi.org/10.1186/s13321-018-0325-4).
18. Norinder U, Boyer S. Binary classification of imbalanced datasets using conformal prediction. J Mol Graph Model 2017;72:256–65. doi:[10.1016/j.jmgm.2017.01.008](https://doi.org/10.1016/j.jmgm.2017.01.008).
19. Toccaceli P. MICP, (2023) (n.d.). <https://github.com/ptocca/>.
20. Alvarsson J, Arvidsson McShane S, Norinder U, Spjuth O. Predicting With confi- dence: using conformal prediction in drug discovery. J Pharm Sci 2021;110:42–9. doi:[10.1016/j.xphs.2020.09.055](https://doi.org/10.1016/j.xphs.2020.09.055).
21. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learn- ing: an introduction to concepts and methods. Mach Learn 2021;110:457–506. doi:[10.1007/s10994-021-05946-3](https://doi.org/10.1007/s10994-021-05946-3).
22. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learn- ing for computer vision? Adv Neural Inf Process Syst 2017:5575–85. <https://arxiv.org/abs/1703.04977> 2017, Decem.
23. Linusson H, Johansson U, Boström H, Löfström T. Reliable confidence predictions using conformal prediction. Lect Notes Comput Sci 2016 (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). doi:[10.1007/978-3-319-31753-3\_7](https://doi.org/10.1007/978-3-319-31753-3_7).
24. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 2021;49:D1388–95. doi:[10.1093/nar/gkaa971](https://doi.org/10.1093/nar/gkaa971).
25. Gaulton A, Hersey A, Nowotka ML, Patricia Bento A, Chambers J, Mendez D, Mu- towo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res 2017;45:D945–54. doi:[10.1093/nar/gkw1074](https://doi.org/10.1093/nar/gkw1074).
26. Sterling T, Irwin JJ. ZINC 15 - ligand discovery for everyone. J Chem Inf Model 2015;55:2324–37. doi:[10.1021/acs.jcim.5b00559](https://doi.org/10.1021/acs.jcim.5b00559).
27. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, en- hanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 2012;55:6582–94. doi:[10.1021/jm300687e](https://doi.org/10.1021/jm300687e).
28. Ursu O, Holmes J, Bologa CG, Yang JJ, Mathias SL, Stathias V, Nguyen DT, Schürer S, Oprea T. DrugCentral 2018: an update. Nucleic Acids Res 2019;47:D963–

70. doi:[10.1093/nar/gky963](https://doi.org/10.1093/nar/gky963).

1. Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, Mathias SL, Nelson SJ, Oprea TI. DrugCentral: online drug compendium. Nucleic Acids Res 2017;45:D932–

9. doi:[10.1093/nar/gkw993](https://doi.org/10.1093/nar/gkw993).

1. Chevillard F, Kolb P. SCUBIDOO: a large yet screenable and easily search- able database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. J Chem Inf Model 2015;55:1824–35. doi:[10.1021/acs.jcim.5b00203](https://doi.org/10.1021/acs.jcim.5b00203).
2. Yang T, Li Z, Chen Y, Feng D, Wang G, Fu Z, Ding X, Tan X, Zhao J, Luo X, Chen K, Jiang H, Zheng M. DrugSpaceX: a large screenable and synthetically tractable database extending drug space. Nucleic Acids Res 2021;49:D1170–8. doi:[10.1093/nar/gkaa920](https://doi.org/10.1093/nar/gkaa920).
3. [Smith L, Gal Y. Understanding measures of uncertainty for adversarial example de- tection. In: Proceedings of the 34th conference on uncertainty in artificial intelli- gence 2018, UAI 2018, 2; 2018. p. 560–9.](http://refhub.elsevier.com/S2667-3185(23)00014-4/sbref0059)
4. Houlsby N, Huszár F, Ghahramani Z, Lengyel M. Bayesian active learning for classification and preference learning, ArXiv. (2011) arXiv ID: 1112.5745. <http://arxiv.org/abs/1112.5745>.
5. Hein M, Andriushchenko M, Bitterwolf J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 41–50. 2019-June. doi:[10.1109/CVPR.2019.00013](https://doi.org/10.1109/CVPR.2019.00013).
6. Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high con- fidence predictions for unrecognizable images. In: Proceedings of the IEEE confer- ence on computer vision and pattern recognition; 2015.
7. Kumar K, Chupakhin V, Vos A, Morrison D, Rassokhin D, Dellwo MJ, Mc- Cormick K, Paternoster E, Ceulemans H, Desjarlais RL. Development and imple- mentation of an enterprise-wide predictive model for early absorption, distribu- tion, metabolism and excretion properties. Future Med Chem 2021;13:1639–54. doi:[10.4155/fmc-2021-0138](https://doi.org/10.4155/fmc-2021-0138).
8. Siblini W, Fréry J, He-Guelton L, Oblé F, Wang Y.Q. Master your metrics with cali- bration, ArXiv. (2019) 457–69. doi:[10.1007/978-3-030-44584-3](https://doi.org/10.1007/978-3-030-44584-3).
9. Morger A, Svensson F, Arvidsson McShane S, Gauraha N, Norinder U, Spjuth O, Volkamer A. Assessing the calibration in toxicological *in vitro* models with conformal prediction. J Cheminform 2021;13:1–14. doi:[10.1186/s13321-021-00511-5](https://doi.org/10.1186/s13321-021-00511-5).
10. Dragos H, Gilles M, Alexandre V. Predicting the predictability: a unified approach to the applicability domain problem of qsar models. J Chem Inf Model 2009;49:1762–

76. doi:[10.1021/ci9000579](https://doi.org/10.1021/ci9000579).

1. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Öberg T, Todeschini R, Fourches D, Varnek A. Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J Chem Inf Model 2008;48:1733–46. doi:[10.1021/ci800151m](https://doi.org/10.1021/ci800151m).
2. Liu R, Wallqvist A. Molecular similarity-based domain applicability metric eﬃ- ciently identifies out-of-domain compounds. J Chem Inf Model 2019;59:181–9. doi:[10.1021/acs.jcim.8b00597](https://doi.org/10.1021/acs.jcim.8b00597).
3. Sheridan RP. Three useful dimensions for domain applicability in QSAR models using random forest. J Chem Inf Model 2012;52:814–23. doi:[10.1021/ci300004n](https://doi.org/10.1021/ci300004n).
4. Sheridan RP. The relative importance of domain applicability metrics for estimat- ing prediction errors in QSAR varies with training set diversity. J Chem Inf Model 2015;55:1098–107. doi:[10.1021/acs.jcim.5b00110](https://doi.org/10.1021/acs.jcim.5b00110).
5. Klingspohn W, Mathea M, Ter Laak A, Heinrich N, Baumann K. Eﬃciency of different measures for defining the applicability domain of classification models. J Chemin- form 2017;9:1–17. doi:[10.1186/s13321-017-0230-2](https://doi.org/10.1186/s13321-017-0230-2).
6. Mathea M, Klingspohn W, Baumann K. Chemoinformatic classification methods and their applicability domain. Mol Inform 2016;35:160–80. doi:[10.1002/minf.201501019](https://doi.org/10.1002/minf.201501019).
7. Janet JP, Duan C, Yang T, Nandy A, Kulik HJ. A quantitative uncertainty metric con- trols error in neural network-driven chemical discovery. Chem Sci 2019;10:7913–22. doi:[10.1039/c9sc02298h](https://doi.org/10.1039/c9sc02298h).
8. Forreryd A, Norinder U, Lindberg T, Lindstedt M. Predicting skin sensitizers with confidence — Using conformal prediction to determine applicability domain of GARD. Toxicol Vitr 2018;48:179–87. doi:[10.1016/j.tiv.2018.01.021](https://doi.org/10.1016/j.tiv.2018.01.021).
9. Norinder U, Rybacka A, Andersson PL. Conformal prediction to define applicability domain – a case study on predicting ER and AR binding. SAR QSAR Environ Res 2016;27:303–16. doi:[10.1080/1062936X.2016.1172665](https://doi.org/10.1080/1062936X.2016.1172665).
10. Jiang Y, Foret P, Yak S, Roy DM, Mobahi H, Dziugaite GK, Bengio S, Gunasekar S, Guyon I, Neyshabur B. NeurIPS 2020 competition: predicting generalization in deep learning, ArXiv. (2020) arXiv ID: 2012.07976. <http://arxiv.org/abs/2012.07976>.
11. Martin CH, Peng TS, Mahoney MW. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nat Commun 2021;12:1–

13. doi:[10.1038/s41467-021-24025-8](https://doi.org/10.1038/s41467-021-24025-8).

1. Rohrer SG, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. J Chem Inf Model 2009;49:169–84. doi:[10.1021/ci8002649](https://doi.org/10.1021/ci8002649).
2. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model 2018;58:916–32. doi:[10.1021/acs.jcim.7b00403](https://doi.org/10.1021/acs.jcim.7b00403).
3. Martin EJ, Zhu XW. Collaborative profile-QSAR: a natural platform for building col- laborative models among competing companies. J Chem Inf Model 2021;61:1603–

16. doi:[10.1021/acs.jcim.0c01342](https://doi.org/10.1021/acs.jcim.0c01342).