Supporting information to MELLODDY: cross pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information

Wouter Heyndrickx,[1] Lewis Mervin,[2] Tobias Morawietz,[3] Noé Sturm,[4] Lukas Friedrich,[5] Adam Zalewski,[6] Anastasia Pentina,[7] Lina Humbeck,[8] Martijn Oldenhof,[9] Ritsuya Niwayama,[21] Peter Schmidtke,[10] Nikolas Fechner,[4] Jaak Simm,[9] Adam Arany,[9] Nicolas Drizard,[11] Rama Jabal,[11] Arina Afanasyeva,[12] Regis Loeb,[9] Shlok Verma,[13] Simon Harnqvist,[13] Matthew Holmes,[13] Balazs Pejo,[14] Maria Telenczuk,[15] Nicholas Holway,[4] Arne Dieckmann,[16] Nicola Rieke,[17] Friederike Zumsande,[6] Djork-Arné Clevert,[7] Michael Krug,[5] Christopher Luscombe,[13] Darren Green,[13] Peter Ertl,[4] Peter Antal,[18] David Marcus,[13] Nicolas Do Huu,[11] Hideyoshi Fuji,[12] Stephen Pickett,[13] Gergely Acs,[14] Eric Boniface,[19] Bernd Beck,[8] Yax Sun,[20] Arnaud Gohier,[21] Friedrich Rippmann,[5] Ola Engkvist,[22] Andreas H. Göller,[3] Yves Moreau,[9] Mathieu N. Galtier,[23] Ansgar Schuffenhauer,[4] Hugo Ceulemans[*1]

[1]Janssen Pharmaceutica NV, Turnhoutseweg 30 Beerse, 2340, BE
[2]AstraZeneca R&D, Biomedical Campus, 1 Francis Crick Ave Cambridge, CB2 0SL, UK
[3]Bayer Pharma AG, Global Drug Discovery, Chemical Research, Computational Chemistry, Aprather Weg 18 a Wuppertal, 42096, DE
[4]Novartis Institutes for BioMedical Research, Novartis Campus Basel, 4002, CH
[5]Merck KGaA, Global Research & Development, Frankfurter Strasse 250 Darmstadt, 64293 , DE
[6]Amgen Research (Munich) GmbH, Staffelseestraße 2 Munich, 81477 , DE
[7]Bayer AG, Machine Learning Research, Research & Development, Pharmaceuticals, Bayer AG, - Berlin, 10117, DE
[8]BI, Medicinal Chemistry Department, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, Biberach an der Riss, 88397, DE
[9]KU Leuven, ESAT-STADIUS, Kasteelpark Arenberg 10, Heverlee, 3001, BE
[10]Discngine, Avenue Ledru Rollin 79 Paris, 75012, FR
[11]Iktos, 65 rue de Prony Paris, 75017, FR
[12]Modality Informatics Group, Digital Research Solutions, Advanced Informatics & Analytics. Astellas Pharma Inc., 21, Miyukigaoka Tsukuba-shi, Ibaraki, 305-8585, JP
[13]GlaxoSmithKline, Computational Sciences, Gunnels Wood Road Stevenage, Herts, SG1 2NY, UK
[14]Budapest University of Technology and Economics, Department of Networked Systems and Services, Műegyetem rkp. 3. Budapest, 1111, HU
[15]Owkin, 12 Rue Martel Paris, 75010, FR
[16]Bayer AG, API Production, Product Supply, Pharmaceuticals, Ernst-Schering-Straße 14 Bergkamen, 59192, DE
[17]NVIDIA GmbH, Floessergasse 2 Munich, 81369, DE
[18]Budapest University of Technology and Economics, Department of Measurement and Information Systems, Műegyetem rkp. 3. Budapest, 1111, HU
[19]Substra Foundation - Labelia Labs, 4 rue Voltaire Nantes, 44000, FR
[20]Amgen Research, 1 Amgen Center Drive Thousand Oaks, CA, 92130, USA
[21]Institut de recherches Servier, 125 chemin de ronde Croissy-sur-Seine, Île-de-France, 78290, FR
[22]AstraZeneca, Molecular AI, Discovery Sciences, R&D, Pepparedsleden 1 Mölndal, 431 50, SE
[23]Owkin, 4 Rue Voltaire Nantes, 44000, FR
[*]Corresponding author email: hceulema@its.jnj.com

Contents

Methods

Data preparation

Private data preparation

Full details can be retrieved from the data preparation manual which is made available as supporting information together with this manuscript. During the project, the manual served as guideline to the private data preparation at the different pharma partners. The end result of that data preparation is a set of files that can serve as input to the machine learning with SparseChem.[1]

Public data preparation

The public data were extracted from ChEMBL (version 25)[2] which provides a curated database of 1.8M drug-like molecules measured on 1.1M assays. Three different assay categories were extracted: ADME and toxicity assays, physical-chemistry assays, and binding and functional assays. Given that the ChEMBL dataset is very heterogeneous, some filters were applied to extract only clean and relevant data. All the following samples were removed: assays without pchembl value, assays with less than 50 points, assays with type not in ['IC50', 'EC50', 'Ki', 'Kd', 'Potency', 'AC50']. The unit was standardized to nM. Then, the physical-chemistry assays have been manually merged when experimental protocols (pH, solvent, etc.) are similar to each other. A cutoff of 8 has been applied on the confidence score provided by ChEMBL which evaluates the relationship between the target and the assay for the binding/functional assays to keep only homologous or direct single protein target.
Compared to private data, public data are a collection of assays measured on a relatively small number of compounds – often from the same chemical series. These data are usually generated and shared by different contributors; therefore, the same assay can be measured differently. To maintain a reasonable number of assays (given the number of compounds) and to keep them 'predictable' i.e. with a sufficient number of measured compounds, a merging strategy has been developed and applied on functional assays similar to each other if they had the same units of measurement and comparable value distributions. The DBScan clustering algorithm was used using the Kolmogorov-Smirnov statistic as distance. Two value distributions have been considered comparable if belonging to the same cluster, the outliers being left alone by the algorithm as expected. This method led to a diminution of approximately 20% of the total number of assays in the output public dataset. The code to generate this data has been made available on GitHub.[3]

High-content-data-based pseudolabels

In the context of modelling pharmacological assays, previous work has pointed towards the value of high content data such as those resulting from cellular imaging experiments.[4,5] Such data can serve as a data source complementary to chemistry-based features such as molecular fingerprints. Federated modelling directly on the cell morphological features extracted from the images was infeasible due to the discrepancies between approaches of the different pharma partners. Instead, high-content-based (image-

based) predictions were generated and transformed into auxiliary tasks which could flexibly be included into the Y matrix without the need to reveal any potentially sensitive information. More concretely, we built a predictive model for all targets in the partner's dataset using the image-based descriptors, and used the predictions (after quality and confidence filtering) as pseudolabels in a set of auxiliary tasks (see Figure S1). The introduction of this type of image data was optional, and one of the advantages of the approach was that partners had the option not to reveal the inclusion of such auxiliary tasks if desired.

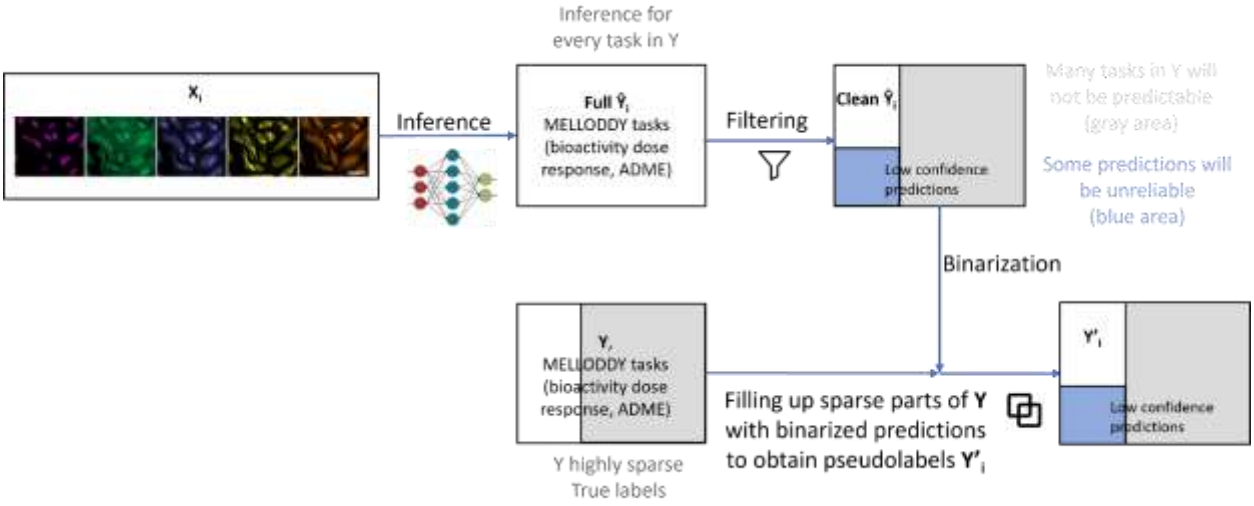


Figure S1. Scheme of the core of the pipeline to generate pseudolabel data. X_i: high content feature input matrix. $\hat{Y}_i$: high-content-based prediction matrix, Y'_i: high-content-based pseudolabel matrix, Y: original label matrix.

The pipeline consisted of three main steps:

1) Training of a model based on high content data
2) Transformation of that model's predictions to binary pseudolabels as auxiliary tasks (Figure S1)
3) Concatenation of the new auxiliary tasks to the MELLODDY dataset

The pseudolabel approach was limited to the classification setting and the resulting AUX_PL tasks served as complement to the AUX_HTS tasks.

The reliability of the predictions from the high content data model were enforced in two ways. As a selection on the columns, an overall task performance of 0.8 negative and positive predictive value (NPV/PPV) was recommended in order for the task to be considered as candidate (quality filtering). As a selection on the rows, only the singleton predictions from the conformal prediction framework were considered, constituting the predictions where the model can confidently assign a single class (confidence filtering). The conformal prediction code from previous work was used,[6] setting the error rate to 0.05.

Enforcing the quality and confidence of the pseudolabels was important since analysis showed that information transfer from the auxiliary tasks to the main tasks would happen even when the quality of the pseudolabels degraded, hereby degrading the predictive performance of the main tasks.

Pseudolabel auxiliary tasks hence mirrored the corresponding main tasks of interest through the lens of the used high content data source. If a main task is predicted effectively by the high content data source, the auxiliary task will consist of a large number of data points, and extensive information transfer to the

main task can be expected (Figure S2). More specifically, this information transfer would especially be expected in unlabeled space since the auxiliary task predictions would inform compounds where the labels are unknown. Established performance metrics such as AUC-ROC or AUC-PR are unsensitive to such changes, therefore the cross-AUC metric is considered, which captures such changes by comparing the auxiliary task pseudolabels with the main task predictions. The cross-AUC-ROC is defined as the AUC-ROC calculated by comparing the compound predictive probabilities for the main task with the 'labels' from the image-based auxiliary task, i.e. the pseudolabels derived from image model predictions. In other words, the cross-AUC can be seen as a measure for how well a main task is predicting its corresponding pseudolabels. So, the AUC is calculated over two tasks, hence its prefix 'cross'. Figure S3 demonstrates two main points: 1) the variation along the x-axis (AUC-ROC) is much smaller than the variation along the y-axis (cross-AUC-ROC), and 2) There is a positive effect upon inclusion of the pseudolabel-based auxiliary tasks.

The full pipeline to generate pseudolabel data has been made available on GitHub.[7]

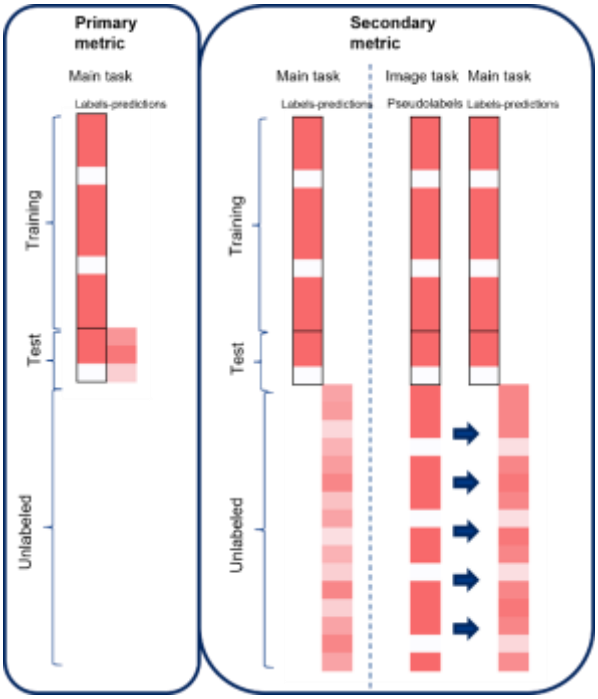


Figure S2 Illustration of information transfer upon inclusion of pseudolabels.In the left pane, the situation without pseudolabels and the performance evaluation through a held-out test set. In the right pane, the performance in unlabeled space underlies the cross-AUC metric. It can be seen as a measure for how well a main task is predicting its corresponding pseudolabels.

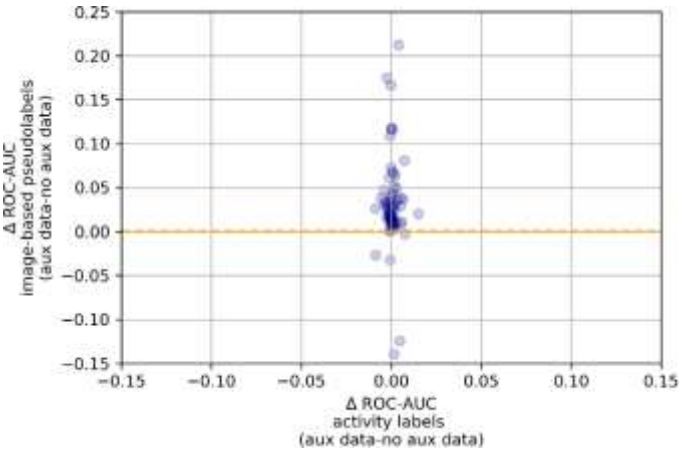


Figure S3. Comparison between the performance on a held-out test set (x-axis, delta AUC-ROC), and the performance in unlabeled space (y-axis, delta cross-AUC-ROC). Blue dots represent task performances.

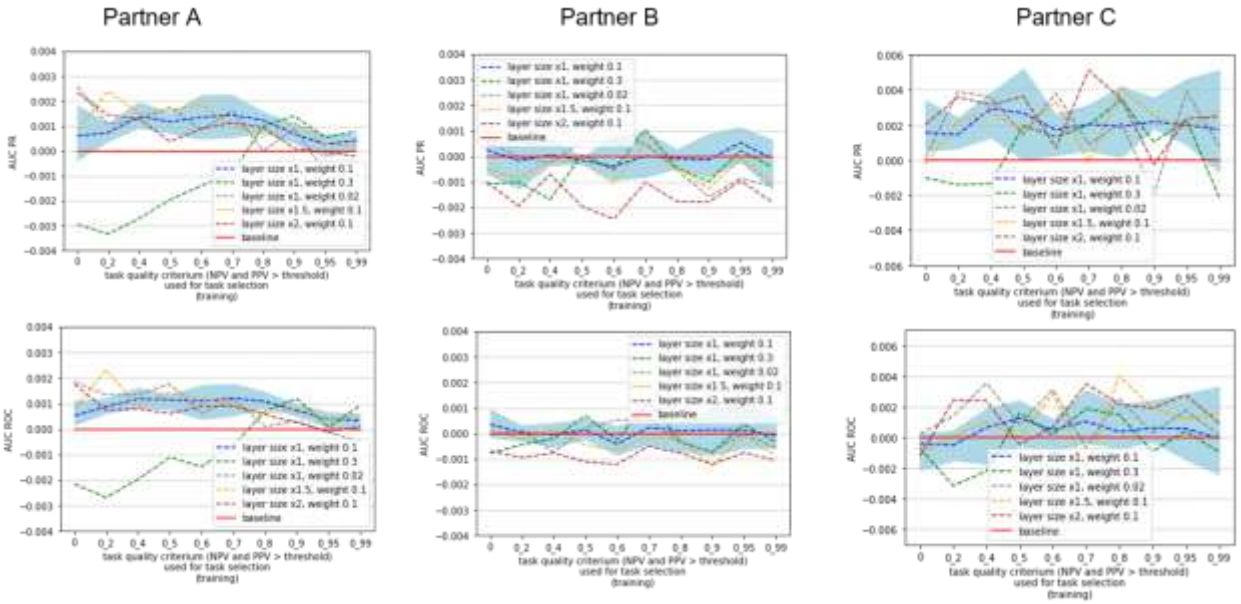


Figure S4. Out of three partners, two report gains in AUC-PR and AUC-ROC upon inclusion of image-based auxiliary tasks. A set of models with different hyperparameters (line plots) or identical hyperparameters (blue envelop) are shown. Results show that intermediate values of quality admission criteria tend to be most favorable, possibly illustrating a trade-off between data volume and conflicting signal.

MELLODDY-TUNER

Following defining the data entry criteria and its subsequent data extraction, a data preparation workflow has been developed and executed on each partners' own environment to ensure consistent standardization of chemical structures and bioactivity across consortium members. An associated open-source python package called MELLODDY-TUNER[8] has been implemented covering chemical structure processing, descriptor calculation, fold assignment, replicate aggregation, and several data quality assessments and filtering (see the data preparation manual for full details). All operations related to

chemical structure processing are built on the open-source cheminformatic software package RDKit.[9] Structure processing includes structure sanity checks, tautomer standardization and calculation of ML-suitable molecular representations, in this case chemical fingerprints (Extended Connectivity Fingerprints (ECFP),[10] implemented as Morgan fingerprints in RDKit), which serve as input features for training a ML model. To obtain uniform assignment of training, validation and test fold across partners without sharing sensitive compound information, a scaffold-based approach has been integrated into the data preparation tool.[11,12] After replicate aggregation, an automated thresholding procedure generates classification tasks and their task weights. Training and evaluation data volume quorum checks are applied to filter both classification and regression tasks. These quora vary given the assay type. Exact configuration settings can be found in the data preparation manual. Finally, matrix representations of the input features (X) and the output tasks (Y) are stored in a format suitable for modelling.

Evaluation

Assay types

Full details on the assignment of assays to assay types can be retrieved from the data preparation manual, particularly in section 3.2. Here we limit ourselves to extending on the manual's contents.

In the manuscript, one subcategory of assays are the alive assays. Alive assays are defined as assays characterized by having at least one datapoint between April 2021 – December 2021, while also meeting the training quora with data prior to April 2021. The period from April 2021 – December 2021 corresponds to the time interval between the two last federated runs, and was chosen for practical reasons.

Another subcategory is the first line safety panel assays. This is the general panel of safety pharmacology assays, that compounds typically get submitted to first before submission to more comprehensive panels at a later stage of the discovery project. Typically, submissions are made jointly to all assays of that panel, resulting in a densely populated block of the Y matrix for these assays. An example is a panel of ~40 assays described in Bowes et al..[13] 25 assays were posed as the minimum volume for the analysis. This panel may show significantly overlap with the assays used in catalogue fusion. Not all partners designated this assay type in their datasets.

While the data preparation manual distinguishes between catalog and non-catalog panel assays, both assay types have been combined in the manuscript, and any aggregations treat both types separately and with equal weight.

Non-evaluated tasks and datapoints

Table S1. Data volume quora which classification (CLS) and regression (REG) tasks had to fulfill to be considered as endpoint, in training, and in evaluation. Only tasks that passed the classification training quorum were considered for regression models.

Assay type	*Quorum Type*	*Classification*	*Regression*
Primary Prediction endpoints *NON-CATALOG* *PANEL* *OTHER* *ADME*	*Training*	*25 observations per class label (active/inactive) in whole dataset*	*50 observations in whole dataset* *25 observations without qualifier in whole dataset*
	Evaluation	*25 observations per class label (active/inactive) in each fold*	*50 observations in each fold* *25 observations without qualifier in each fold* *For non-ADME: Standard deviation > 0.5 in all folds*
Primary Prediction endpoints *CATALOG-PANEL*	*Training*	*400 observations in total*	*Not applicable (catalogue fusion is not used for regression) CATALOG-PANEL tasks are treated as NON-CATALOG-PANEL tasks*
	Evaluation	*25 observations per class label (active/inactive) in each fold*	*Not applicable (catalogue fusion is not used for regression) CATALOG-PANEL tasks are treated as NON-CATALOG-PANEL tasks.*
AUX_HTS *AUX_PL*	*Training*	*10,000 measurements* *10 actives*	*Not applicable (no auxiliary data in regression for year 3)*
	Evaluation	*Not applicable*	*Not applicable*

Applicability domain

The conformal efficiency as a metric for a classifier's applicability domain (AD) was calculated by the Mondrian Inductive Conformal Predictor approach,[23,26,41] with the non-conformity function based on the predictive probabilities and an error rate (ε) set to 0.05. Only non-auxiliary tasks of sufficient size, class balance, and predictive quality were considered. Concretely, 25 actives and 25 inactives, and an AUC-ROC of 0.6 was required on the calibration set, which was formed by splitting the original test set in approximately half. Unless stated otherwise, the reported conformal efficiency was calculated on an unlabeled dataset with compounds from an undisclosed commercial catalog, due to its relevance to the use case of finding new hit material with the federated models. This is a dataset consisting of 10,000 randomly sampled compounds from an external commercial, small-molecule drug discovery catalog (commercial catalog 2).[23]

Training details

Phased approach

The training process was organized using the standard training/validation/testing split in the compound

space.[14] More specifically, in Phase 1 hyperparameter tuning of the model was performed using 60% of the data (i.e. 3 out 5 folds) for training and 1 fold for evaluation. In Phase 2 models were trained on 80% of the data (3 training folds from Phase 1 + the validation fold) based on the best hyperparameters identified in Phase 1 and evaluated on the remaining (fifth) test fold. These results are reported in this manuscript. Phase 3 models were trained on all data, and were of practical relevance through internal use by the pharma partners. Due to the lack of data for testing, the performance of the latter models cannot be evaluated within the scope of this project. More details on the procedure and implementation can be found in Oldenhof et al..[15]

Optimal hyperparameters

Table S2. Specifications and ranges of optimal architectures in single-partner models. 'N/A' indicates that the corresponding parameter was not used. Square brackets indicate multi-value hyperparameters. Semicolons separate discrete hyperparameter sets for partners.

	CLS	CLSAUX	REG	HYB
Number of layers	1;2	1;2	1;2	2
Number of epochs	20	20	20	20
Learning rate steps	10	10	[2 5]; [2 10]; [2 15]	[2 5]; [2 10]; [2 15]
Regression weight	N/A	N/A	N/A	0.5;0.75
Dropout rate trunk	0.6 - 0.8	0.4 - 0.8	0.8	0.8
Dropout rate head REG	N/A	N/A	N/A; 0.4	0.4
Dropout rate head CLS	N/A; 0.6	N/A; 0.4 - 0.6	N/A	N/A
Trunk layer size	2000 - 5000	1600 - 12000	1000 - 3000	1000 - 3000
Head layer size REG	N/A	N/A	N/A; 3000 - 5000	1000 - 5000
Head layer size CLS	N/A; 2000	N/A; 4000 - 6000	N/A	N/A

Table S3. specifications and ranges of optimal architectures in multi-partner models. 'N/A' indicates that the corresponding parameter was not used. Square brackets indicate multi-value hyperparameters. Semicolons separate discrete hyperparameter sets for partners.

	CLS	CLSAUX	REG	HYB
Number of layers	1-2	2-3	2	2
Number of epochs	30	50	30; 50	50
Learning rate steps	[10]	[10 30]; [20 30]; [30 40]	[2 10]; [2 8 14]	[2 10]
Regression weight	N/A	N/A	N/A	0.5; 0.75
Dropout rate trunk	0.8	0.6; 0.8	0.8	0.8
Dropout rate head REG	N/A	N/A	0.4; 0.6	0.4; 0.6
Dropout rate head CLS	N/A; 0.4	0.4	N/A	N/A
Trunk layer size	8000;11000	8000; 11000	2000; 6000	2000; 4000; 6000
Head layer size REG	N/A	N/A	8000	6000; 8000
Head layer size CLS	8000; N/A	8000; 11000; [11000, 8000]	N/A	N/A
Number of minibatches	50	80	50	50
Maximum size of internal batch	1000	1000	1000	1000

Hybrid models

The current section discusses the impact of classification auxiliary tasks on regression models and vice versa. The joint modelling of classification and regression tasks results in hybrid models. Here, three hybrid model setups with a different emphasis on regression and classification tasks, were evaluated.

In Figure S5, panel a shows Pareto frontiers of performance differences for the regression and classification metrics between the results of a set of hybrid models and optimal baseline models purely trained on regression and classification tasks, respectively. The contributions of regression and classification tasks during model training were balanced by the regression weight parameter, w_{REG}, which linearly scaled the loss function impact from a pure classification model (w_{REG}=0.0), to a CLS-emphasized hybrid (w_{REG}=0.25), to a balanced hybrid (w_{REG}=0.5), to a REG-emphasized hybrid (w_{REG}=0.75), to a pure regression model (w_{REG}=1.0).

For each value of w_{REG}, Pareto frontiers were extracted from a set of 25 hybrid models trained in Phase 1 on the single-partner level. Panel b shows an identical analysis for another regression metric (correlation coefficient). Results are illustratively shown for one pharma partner. Trends were found to be consistent across multiple partners participating in the investigation. The results of this analysis demonstrate that only regression models benefited from a hybrid model setup which was the motivation for taking a regression-focused hybrid setup to the federated platform in which classification tasks were considered purely as auxiliary data.

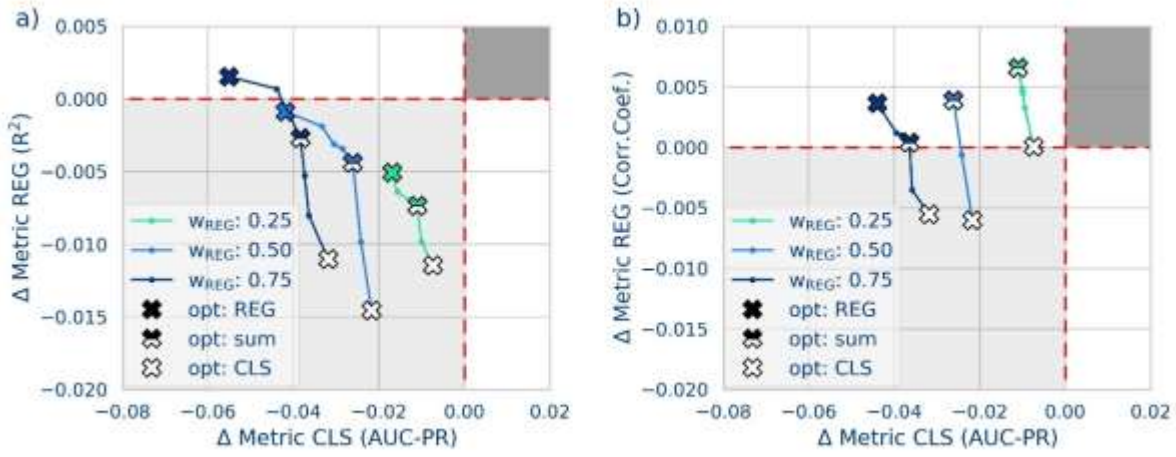


Figure S5. Pareto-plot showing classification and regression model performance vs. the baseline. Three regression weights (w_{REG}) were investigated for a single partner on a full scale multitask dataset, with hyperparameter optimization.

Privacy attacks

Since the code of the clients and the server of the MELLODDY platform is audited and verified,[15] the

privacy analysis only considered passive attacks, specifically membership inference attacks, which infer if a specific record was part of the target model's training dataset.[16] A comprehensive list of other attacks is provided in Liu et al..[17] Within MELLODDY, secure aggregation is applied to prevent attribution, which restricts the access to individual model updates.[18]

Relation to existing work

MELLODDY applied a federated multitask learning conceptually similar to the approach of Smith et al..[19] An example from the latter work, is the modelling of sitting behavior (binary classification of whether or not the user is sitting down) for mobile phone users with features obtained from the phone's accelerometer and gyroscope. Here, a single task was modelled for each user, and the multitask aspect applied to the federated setting where the modelling of the users was done jointly. In this case, multitask learning exploited commonalities between users for predictive performance improvements.

This maps to the MELLODDY case since a partner's tasks (assays) were individual, but related tasks would be present across partners. For all partners, most assays would measure either the binding interaction of small molecules with large target biomolecules, especially proteins, either directly or by detecting downstream effects resulting from such an interaction. To some extent these targets were expected to overlap exactly between partners,[20] although no explicit target mapping has been performed in MELLODDY and the modelling was unaware of the targets associated with the assays. In addition, the partners also shared common pharmacokinetic and physical properties as modelling tasks. Just as modelling sitting behavior for two different people could benefit from joint modelling, modelling the activity on common pharmacological target types or physical properties could similarly benefit across pharma partners.

In the classification framework of Yang et al.,[21] the modelling of sitting behavior corresponds to horizontal federated learning since $X_i = X_j$, $Y_i = Y_j$, $I_i \neq I_j$, with feature space X, label space Y, sample ID space I and i,j indexing different partners. Even though one could argue that the label space is different since we are dealing with individual models, sitting conceptually means the same across users and the users could also be modelled together. Therefore, the label space can be considered the same. Since this is not a priori the case in MELLODDY, we classify MELLODDY as $X_i = X_j$, $Y_i \neq Y_j$, $I_i \neq I_j$.[22] It can be remarked that the catalog assay fusion approach described in this work can be classified as horizontal federated learning, i.e. $X_i = X_j$, $Y_i = Y_j$, $I_i \neq I_j$.

Safety comment

Since the study was limited to computational work, and did not involve lab activities or experiments, no unexpected or unusually high safety hazards were encountered.

Released software

The following software packages have been made available:

- MELLODDY-TUNER
  https://github.com/melloddy/MELLODDY-TUNER
- Model predictive performance evaluation
  https://github.com/melloddy/performance_evaluation
- Pseudolabel data preparation
  https://github.com/melloddy/pseudolabel_auxdata
- Scaffold-based multi-party simulation split
  https://github.com/melloddy/simulated-multi-partner-splits
- Public data preparation
  https://github.com/melloddy/public_data_extraction
- Model manipulation software
  https://github.com/melloddy/MELLODDY-Predictor

Avenues to explore

Model fusion

Given the wealth of obtained models (e.g., arising from different architectures or auxiliary data), exploration of model ensembles as sources of performance gains becomes possible. Such approaches have been shown successful[23] and make particular sense for MELLODDY which optimizes average performance, potentially disregarding optimal models for specific tasks or groups. Still, care is needed to avoid overfitting given the number of available models. To this end, rigorous selection of models for each task based on the validation sets and confirming the benefit on held-out test sets is performed. This also allows evaluating combinations of single- and multi-partner models. Preliminary results obtained from several companies indicated that per-task selection can favor models different than the average top-performers. Limited project time and complexity of designing model groups prevented finding ensembles that would consistently outperform top multi-partner models on test sets but, given the post-processing nature of the analyzes, further selection schemes can still be explored.

Catalog assay fusion

Catalog assay fusion consists of cross-compound federation through assay reconciliation. On the federated platform, this is achieved through a shared catalogue head (Figure S6). The reconcilable CATALOG-PANEL assays were derived from list of safety panel assays, based on identical protocols since they were outsourced to contract research organizations.

To explore the impact of catalog assay fusion on the model performance, an experiment was set up with three virtual partners using public dummy data. The virtual partners were created with the splitting code (ref), based on a random selection of assays from the public dataset (15 assays resulting in 39 tasks). Three

setup were explored: 'fusion' (local, single-partner catalog assay fusion), 'no fusion' (local, single-partner catalog assay fusion disabled) and 'MELLODDY' (federated, multi-partner catalog assay fusion for the tasks of assay type 'CATALOG-PANEL').

A performance increase was observed for the virtual partners comparing catalog fusion enabled ('Fusion') compared to catalog fusion disabled ('No fusion'), both relative to the respective single partner performance (median RIPtoP: Fusion= 22.7% > No fusion = 17.4%). Unfortunately, the promising single-partner results upon cross-compound federation did not materialize on the federated platform (median RIPtoP: MELLODDY = -12.6%) (Figure S6).

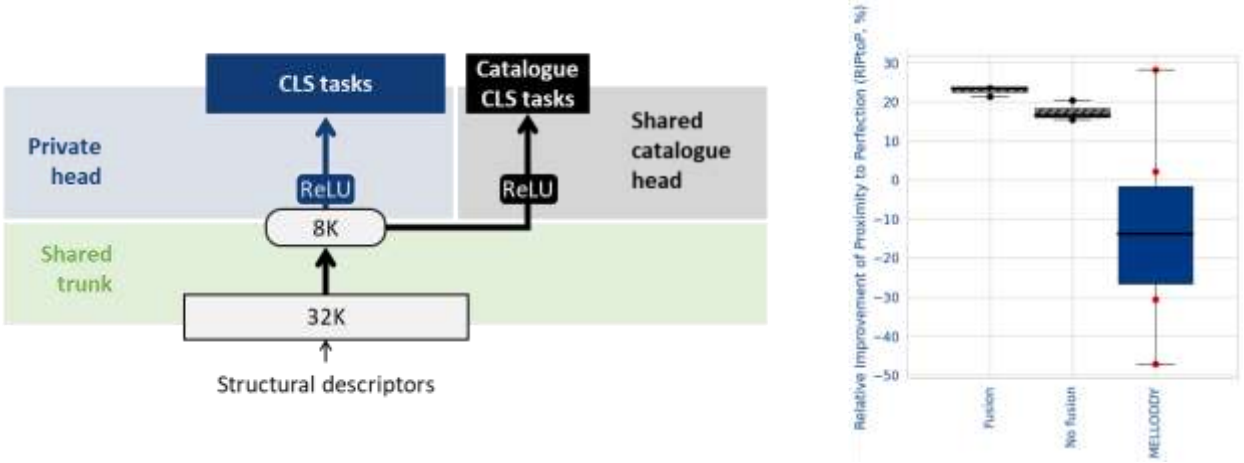


Figure S6. (Left) Catalog fusion scheme. Besides the private head containing all classification tasks, a shared catalogue head is built by mapping common assays across partners to the respective shared tasks. This shared head enables a cross-compound federation of these assays. (Right) Relative improvement of proximity to perfection (RIPtoP, %) for three different settings, showing that the promising single-partner results upon vertical federation (fusion vs. no fusion) did not materialize on the federated platform (MELLODDY).

Local trunk

Another aspect of the implemented algorithm is that it allows for practical postprocessing of the models to allow transfer learning.[24–26] Indeed, the shared trunk of the federated models embeds a unique latent representation as it has absorbed the data of multiple companies. As such, this fixed shared trunk could be used as a molecular representation tool (SMILES-to-MELLODDY embeddings) and further employed in data science applications as input descriptor for other ML tools, that may ingest any other type of potentially complementary compound descriptors, like CDDD.[27] Hereby, the knowledge transfer benefits from the federated learning can be combined with the flexibility of a partner-specific data representation accounting for specifics of a pharma partner's assays and compounds. This approach allows leveraging federated information when locally modelling novel or enlarged assays. The shared trunk brings the benefit of information transfer between partners, but it might overly restrict the learned data representation. This possibility was assessed by complementing the fixed multi-partner trunk with a partner-specific, local trunk, trained on a partner's data. However, experimental results (Figure S7) based on per-partner best multi-partner and single-partner classification models did not demonstrate consistent performance gains.

Training details:

The hyperparameters of the local trunk were selected using the trunk from a multi-partner Phase 1 model, i.e., the best multi-partner model trained on 3 out of 5 folds (i.e. excluding the validation and the test folds). The size of the local trunk size was varied (single-layer options are [200, 400, 800, 1200, 1600, 2000, 2500, 3000], double-layer – [(1200,1200), (1600,1600)]). For dropout, the following search space was used [0.4, 0.5, 0.6, 0.7, 0.8]. The size of the hidden layer in the head was set to 8000.

Based on AUC PR of the validation fold, the best hyperparameter configuration, as well as the best epoch were selected and applied to training a Phase 2 local trunk model using 4 out of 5 folds (excluding the test one) using shared trunk from a multi-partner Phase 2 model trained on the same 4 folds.

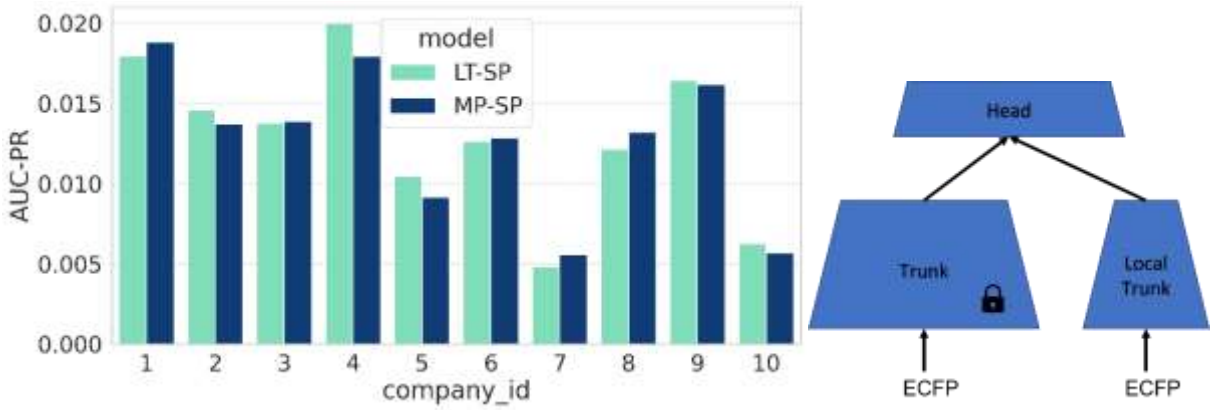


Figure S7. Left: Delta in performance measured by AUC-PR on the test fold for Local Trunk (LT) and multi-partner (MP) classification models, with respect to single-partner (SP) models, demonstrated per pharma partner. Right: schematic illustration of local trunk architecture.

Time-dependent applicability domain

One consequence of an applicability domain extension for the multi-partner models would be a more stable prediction reliability over time compared to single-partner models. Generally, a decline of reliability over time is expected because with more time having passed since training, new compounds are assumed to be more distant to the training data (a similar motivation is given by Sheridan et al.[28] for time-split cross-validation and the over-time-decline observation is also discussed in Davis et al.[29]). This led to the idea for evaluating the applicability domain difference between single- and multi-partner models by quantifying their difference in performance decay over time. For this purpose, a new evaluation metric and algorithm was developed. Unfortunately, the nature of this idea requires having a sufficient timeframe after training passed over which the performance decay can be evaluated and compared. This makes such an evaluation for the final models built with this collaboration infeasible at the time of writing because not enough time has passed since their training data was compiled.

Preliminary experiments run by some partners using models and data from the previous run revealed cases where, depending on a given task, either single- or multi-partner models showed a slower decline, but also indicated challenges for reporting quantitative results. These are mainly due to two limitations: first, the comparatively small amount of data (a specific, limited timeframe and additionally partitioned into time bins) limits the expressiveness of statistical analyses of high-level aggregations. Second, the possibility to report details is very restricted due to the sensitivity of the data, which could indicate assay usage rates over time if shown in detail.

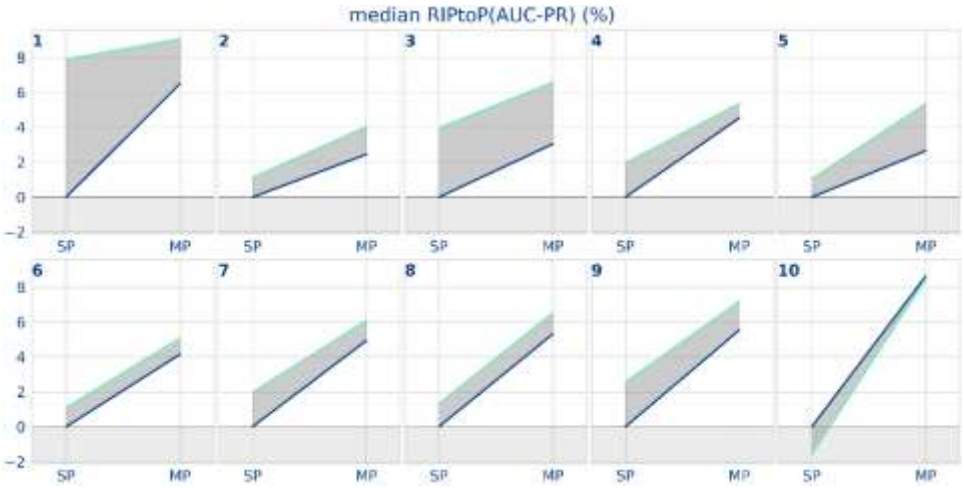
Concluding, while such an analysis provides some interesting insights to the participating companies, based on the currently available data it would be premature to draw conclusions. Consequently, sharing conclusions, results and more details about this approach has to be referred to future work hoping that more available prospective data will allow for that.

Results

Alternative visualizations

In this section some alternative visualizations of the main results are presented, using RIPtoP, which allows for direct comparison with the manuscript.

Figure S8 shows plots which outline the performance delta effects for both inclusion of auxiliary and other partners' data. Performance deltas are presented on a per-partner basis as a function of the performance improvement relative to the single-partner, no auxiliary data, baseline. These plots can be seen as an alternative visualization for the Figures 3A, 4A, 5A in the manuscript. They aim to demonstrate whether the large volume of auxiliary data models increases overlap between the partners and allows the federated learning to extend its gains relative to not using auxiliary data. This would be the case when the SP*/MP* slope in green is larger than the SP/MP slope in blue. This does not consistently seem to be the case.



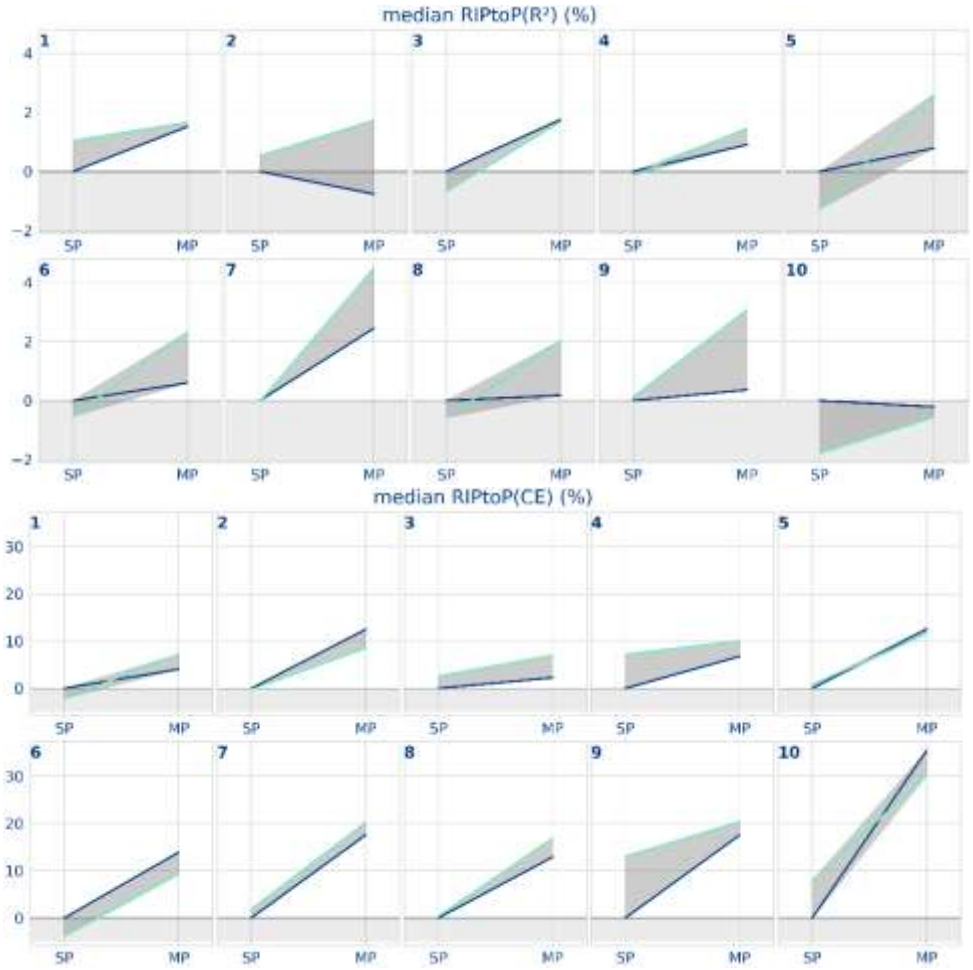


Figure S8. Plots outlining the performance delta effects for both inclusion of auxiliary (*) and other partners' data (MP).

Figure S9, Figure S10 and Figure S11 provide more detail to the ECDF Figures 3C-F, 4C-F and 5C-F in the main paper. Specifically, the maximum slope and the y=0.5 intersection point in the figures indicate that the SP baseline AUC-PR values typically ranged between 0.6 and 0.8 on a scale of 0.0 to 1.0, while average R^2 values were at lower values (around 0.3 to 0.4).

The CDF expresses the fraction of the population with a value lower than the threshold. For example, at AUC ROC = 0.8 the SP curve is somewhat higher than the MP curve. This indicates that there is a higher proportion of SP tasks that have an AUC ROC value lower than 0.8 compared to the MP tasks. Hence, MP is outperforming SP. At AUC ROC = 0.5 there are (close to) zero tasks that have an even lower AUC ROC value. On the other hand, no tasks can exceed the theoretical maximum of AUC ROC = 1. The curves actually split the full set of tasks into two parts at every possible threshold value: one passing the threshold (success, high performance, area above the curve), one not passing the threshold (failure, low performance, area below the curve).

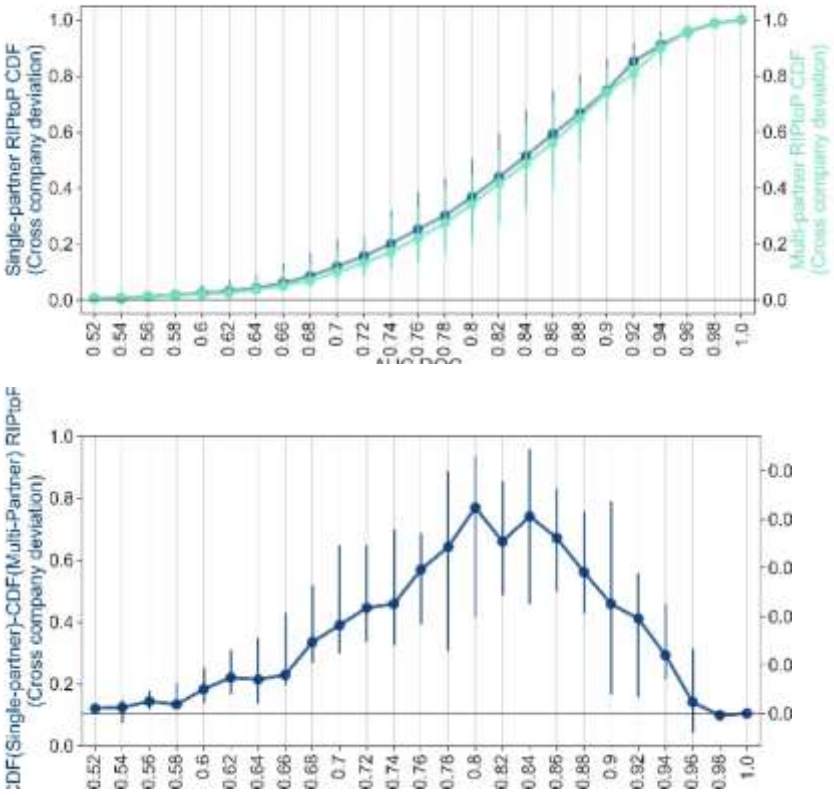


Figure S9. Empirical cumulative distribution plots (CDF) for multi-partner and single-partner models for RIPtoP(AUC-ROC) (top), and the difference between both (bottom). The error bars indicate the interquartile ranges over partners.

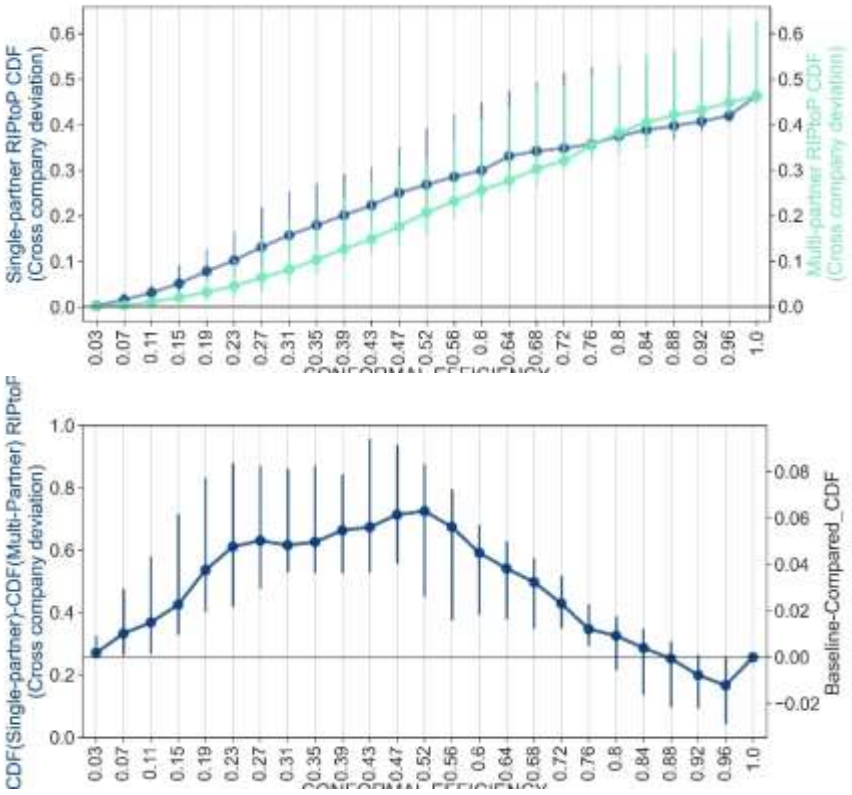


Figure S10. Empirical cumulative distribution plots (CDF) for multi-partner and single-partner models for RIPtoP(CE) (top), and the difference between both (bottom). The error bars indicate the interquartile ranges over partners.

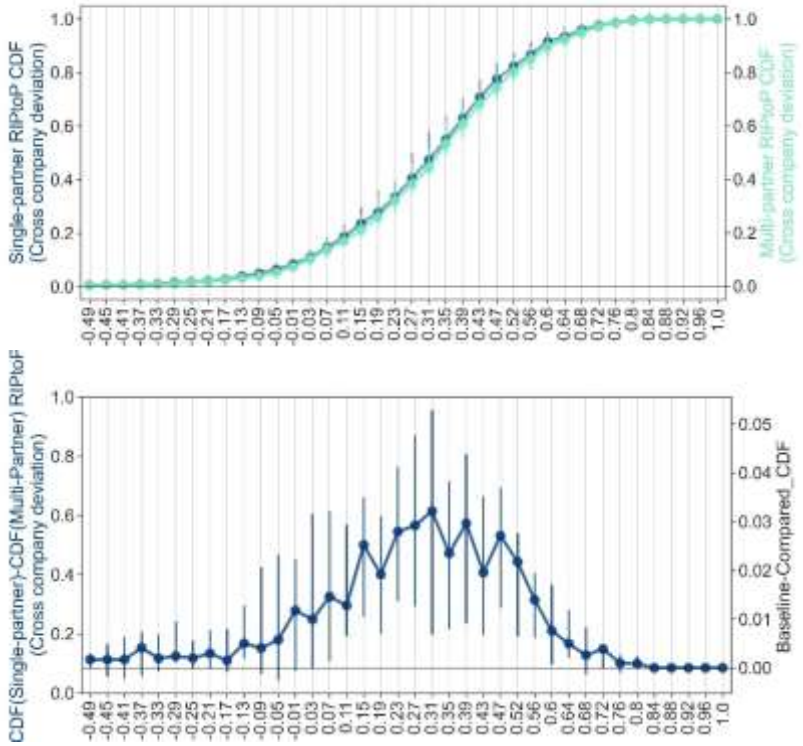


Figure S11. Empirical cumulative distribution plots (CDF) for multi-partner and single-partner models for RIPtoP(R^2) (top), and the difference between both (bottom). The error bars indicate the interquartile ranges over partners.

Applicability domain

The dependency of the ΔCE on the type of unlabeled dataset was explored previously.[6] In this work, summarizing, similar trends between datasets for the final MELLODDY models are found, e.g. DrugSpaceX having the lowest median for 10/10 partners (Figure S12).

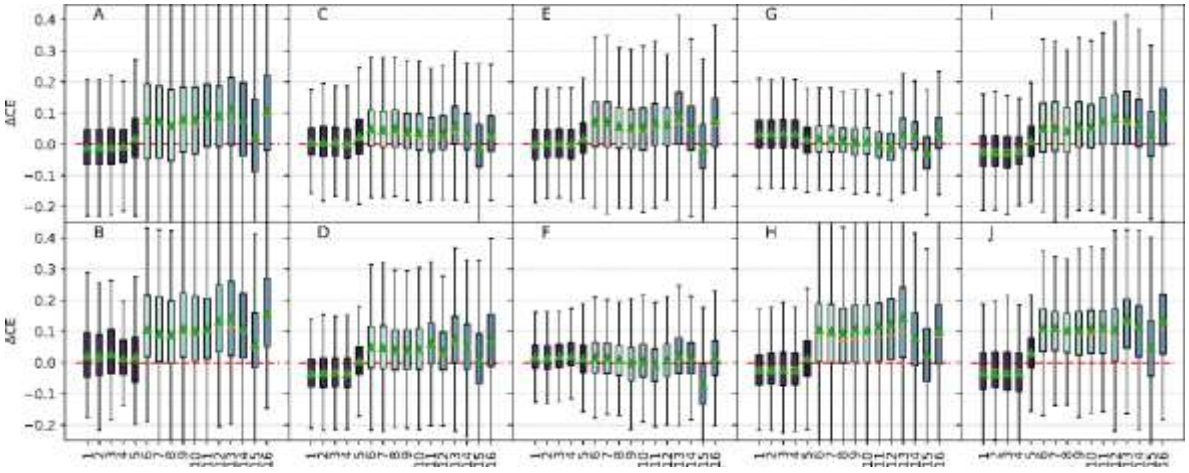


Figure S12. Partner-level predictive confidence analysis via the task-distribution of conformal efficiency difference (ΔCE, MP-SP*) considering several datasets. Lighter shades represent the unlabeled datasets, darker shades the labeled datasets. The different*

datasetes are indexed by integers on the x-axis. DrugSpaceX is indexed by 15. See Heyndrickx et al.[6] for full details on the datasets). Letters A-J correspond to partners and do not necessarily match previous assignment.

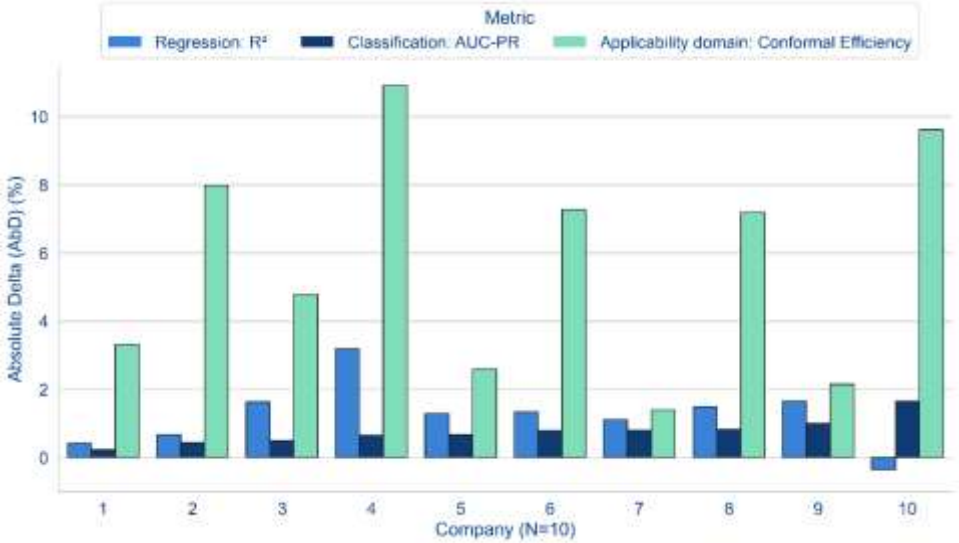
Alternative delta modalities

In this section, figures on alternative delta modalities (i.e. alternatives to the RIPtoP) such as relative improvement (RI) and the absolute delta (Δ) are explored, for classification (including the applicability domain) and regression. These describe task-level performance by comparing the model of interest (MoI) with the baseline.

$$RI(metric) = \frac{metric_{MoI} - metric_{baseline}}{metric_{baseline}} \quad (1)$$

$$\Delta(metric) = metric_{MoI} - metric_{baseline} \quad (2)$$

Figure S13, can be seen as equivalent to the Figure 3 in the main paper. Figure S14, Figure S15, Figure S16 and Figure S17 can be seen as equivalents to the Figures 4A-B, 5A-B and 6A-B in the main paper. They aim to provide more insight into the effect of swapping RIPtoP with another delta modality such as RI or Δ.



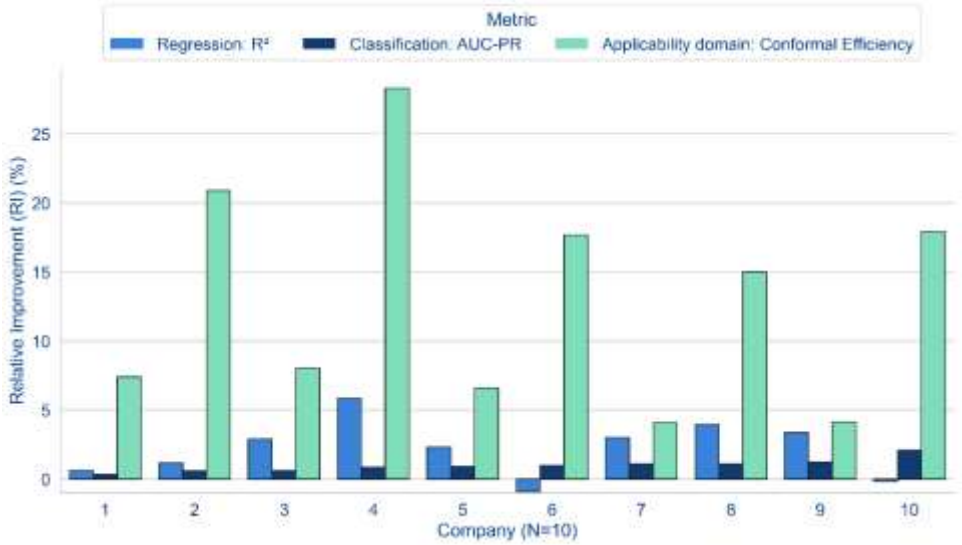


Figure S13. Performance deltas (between multi-and single-partner runs) across-companies for their respective optimal model (i.e., with/without auxiliary data).

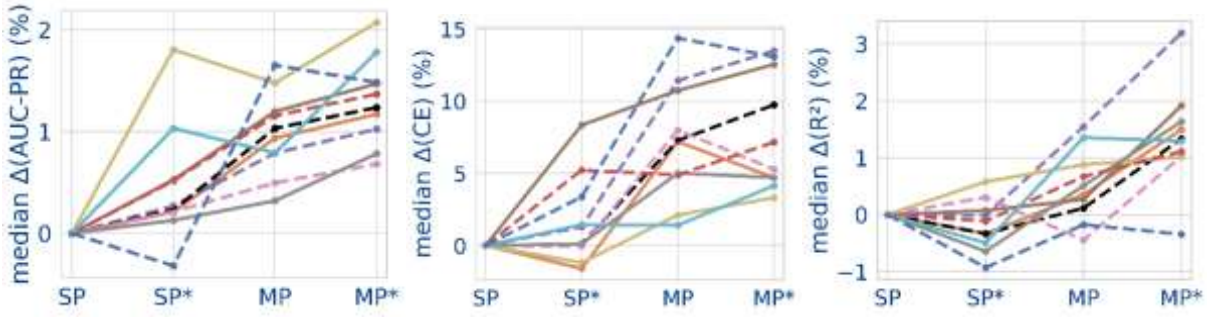


Figure S14. The effect of multi-partner (MP) and auxiliary data (*) on the median task performance, for 5 smaller (dashed lines) and 5 larger (solid lines) partners, for the absolute delta (Δ).

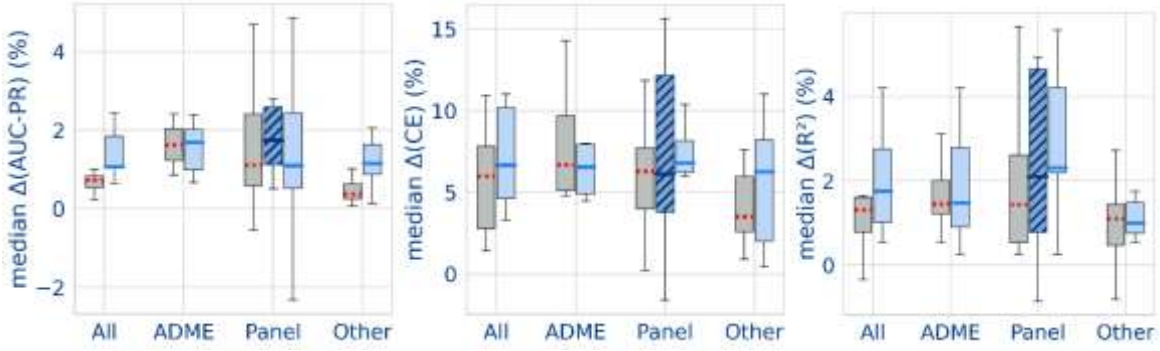


Figure S15. Distribution of median task performance (Δ) over partners.

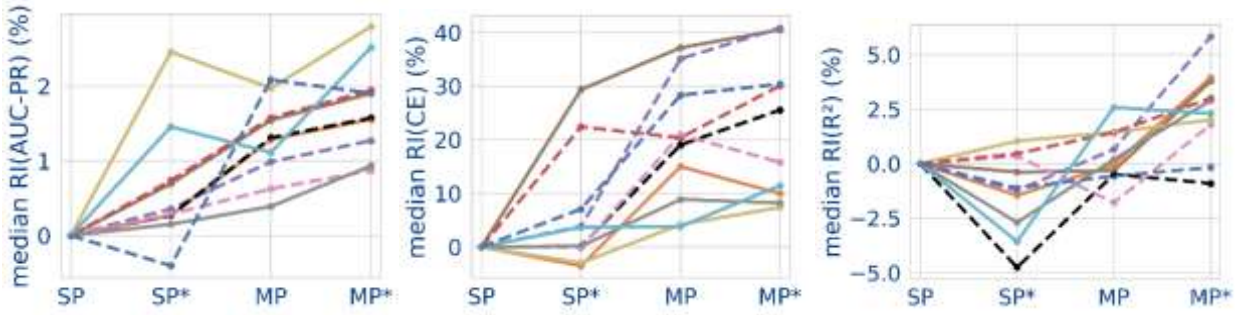


Figure S16. The effect of multi-partner (MP) and auxiliary data (*) on the median task performance, for 5 smaller (dashed lines) and 5 larger (solid lines) partners, for the relative improvement (RI).

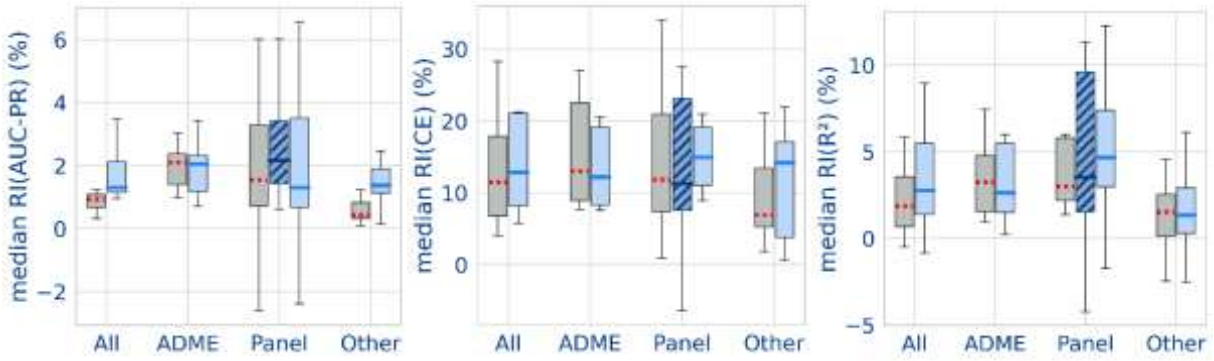
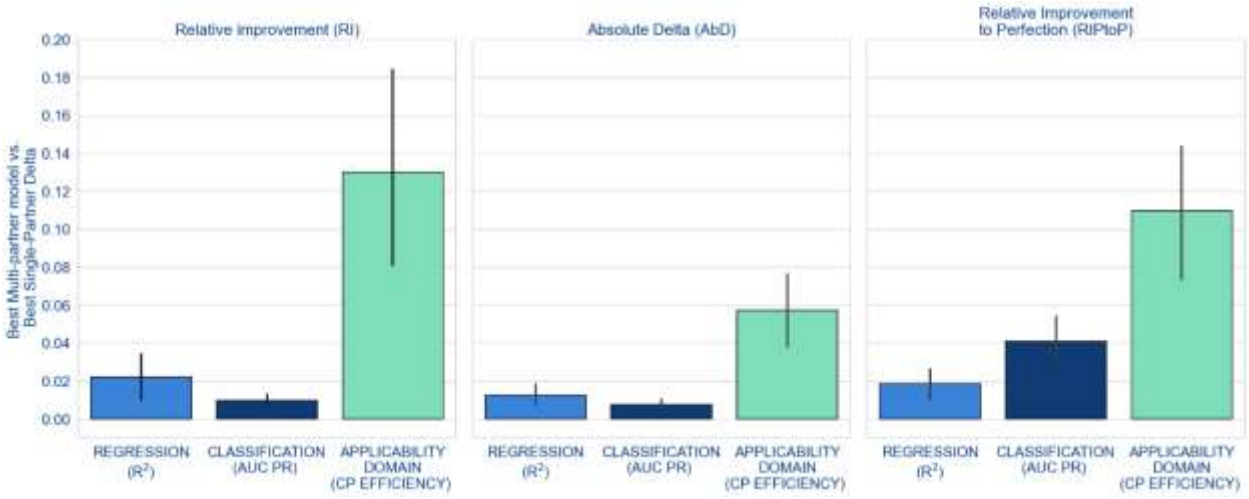


Figure S17. Distribution of median task performance (RI) over partners.



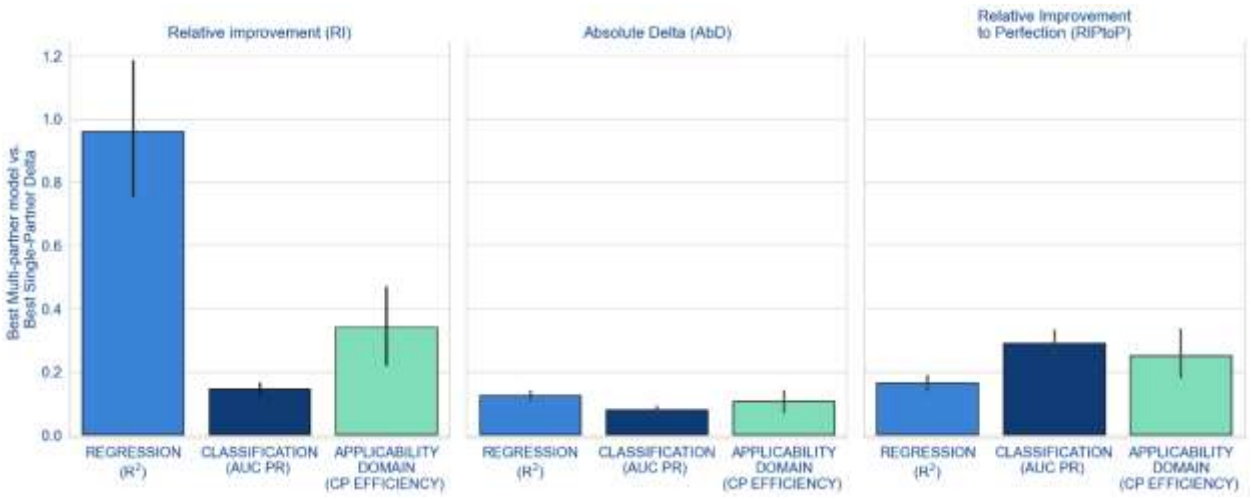


Figure S18: Multi-partner gains averaged over all partners for different absolute and relative delta modalities, i.e., from left to right, relative improvement (RI), absolute delta (Δ), and relative improvement of proximity to perfection (RIPtoP). The top pane shows the median delta modality, the bottom pane the 90th percentile delta modality. The error bars indicate the interquartile ranges over partners.

Figure S19 shows the dependency of the RIPtoP and the RI on the SP baseline. Lower values of the SP baseline lead to higher RI values upon MP improvement (left), while higher values of the SP baseline lead to lower RI values upon MP improvement (right). In other words, relative performance improvements closer to the end of the scale are emphasized by the RIPtoP metric, hereby reflecting the increasing difficulty to improve an already good baseline, which was of higher interest than increasing a poor baseline by a modest margin.

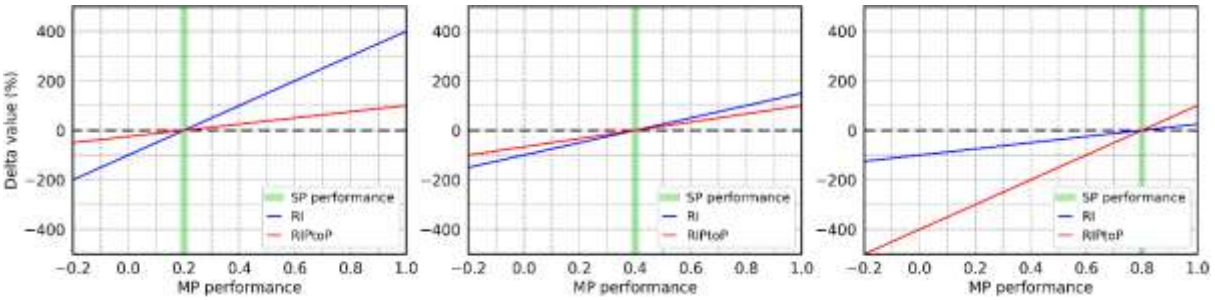


Figure S19: Comparison of two approaches for measuring relative performance improvements, proximity-to-perfection (RIPtoP) (red line) and relative improvement (RI) (blue line), and their dependence on the position of the single-partner baseline value (green bar).

Alternative performance metrics

A commonly used metric as alternative to R^2 for regression is the correlation coefficient. From Figure S20 it can be seen that gains on the correlation coefficient regression metric are positive for all partners.

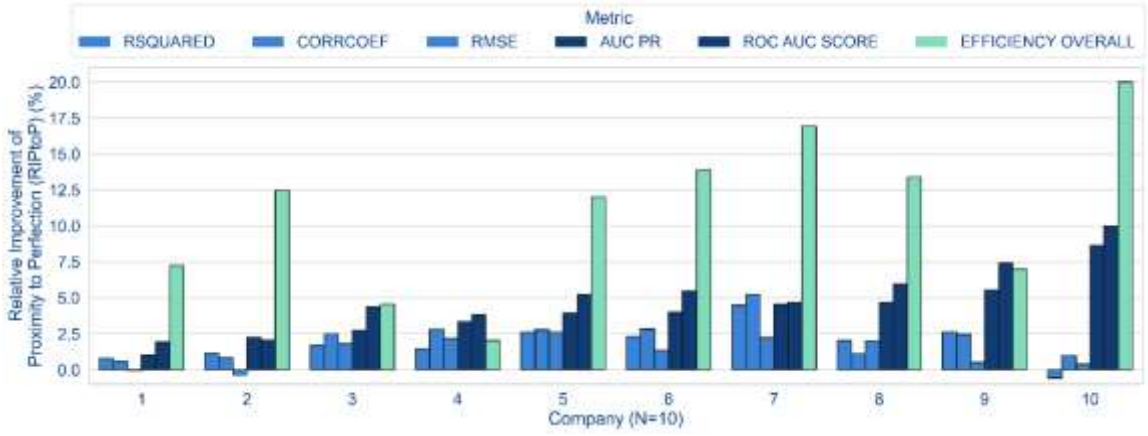


Figure S20. TOC-like figure for showing that the correlation coefficient regression metric are positive for all partners

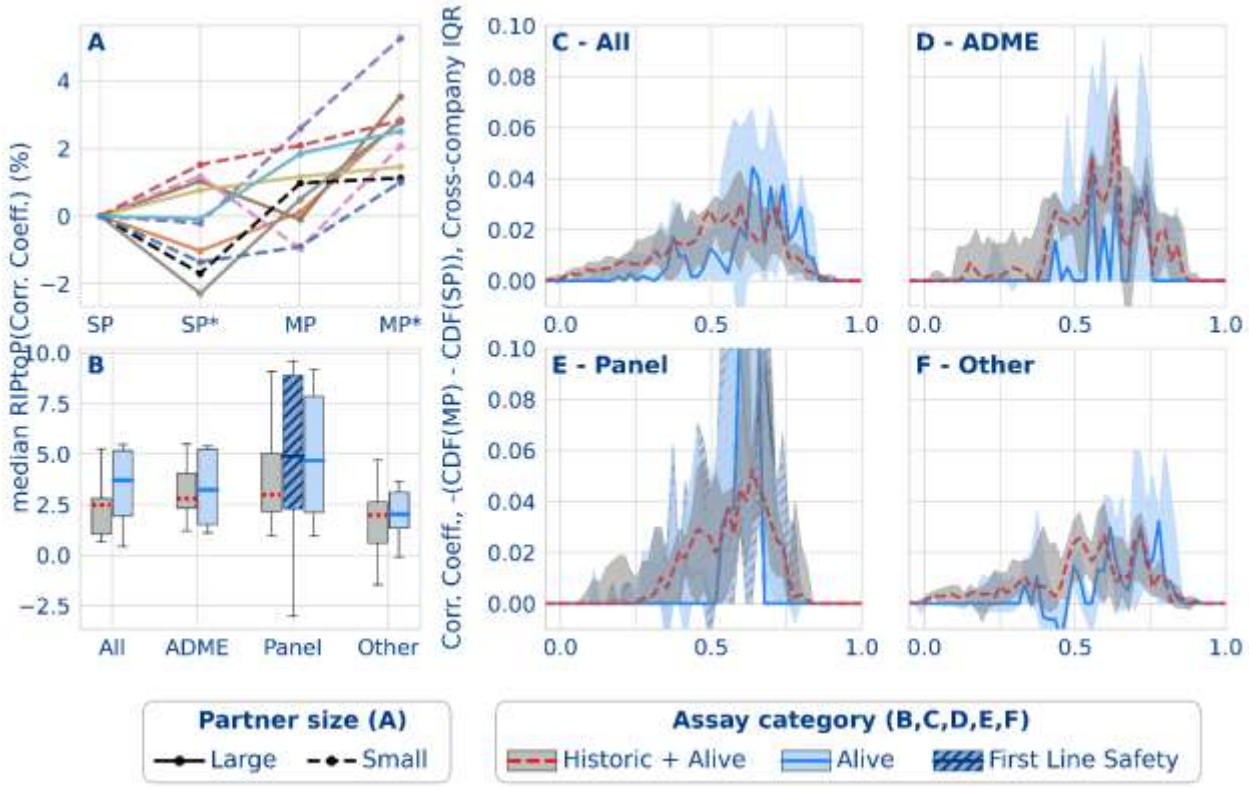


Figure S21. Regression performance results from the federated run. (A) The effect of multi-partner (MP) and auxiliary data (*) on the median task performance, for 5 smaller (dashed lines) and 5 larger (solid lines) partners. (B) Distribution of median task performance RIPtoP(Correlation Coefficient)) over partners. (C-F) Difference between the empirical cumulative distribution functions (CDF) from single and multi-partner models for different assay types based on Correlation Coefficient. The y-axis represents the difference between the proportion of tasks in the multi- versus single-partner models, as a function of a given cumulative performance on the x-axis. The line plots indicate the median probability difference for a bin over partners. The interquartile ranges are indicated with the shaded envelope.

Task size effect

Figure S22 and Figure S23 show the relationship between the task size and the change in task's performance for three partners. The fitted second-order polynomial is shown in black, indicating that there is no strong relationship between the task size and the observed task performance change. For classification, an upward trend can be seen towards larger task sizes, especially for the RIPtoP(AUC-PR). This indicates that the largest tasks benefit the most from federated learning. However, it should be noted

that the relative density of tasks in these large size regions is lower than average, and as such, the upwards trend was caused by a relatively small number of tasks. No such trend could be observed for regression.

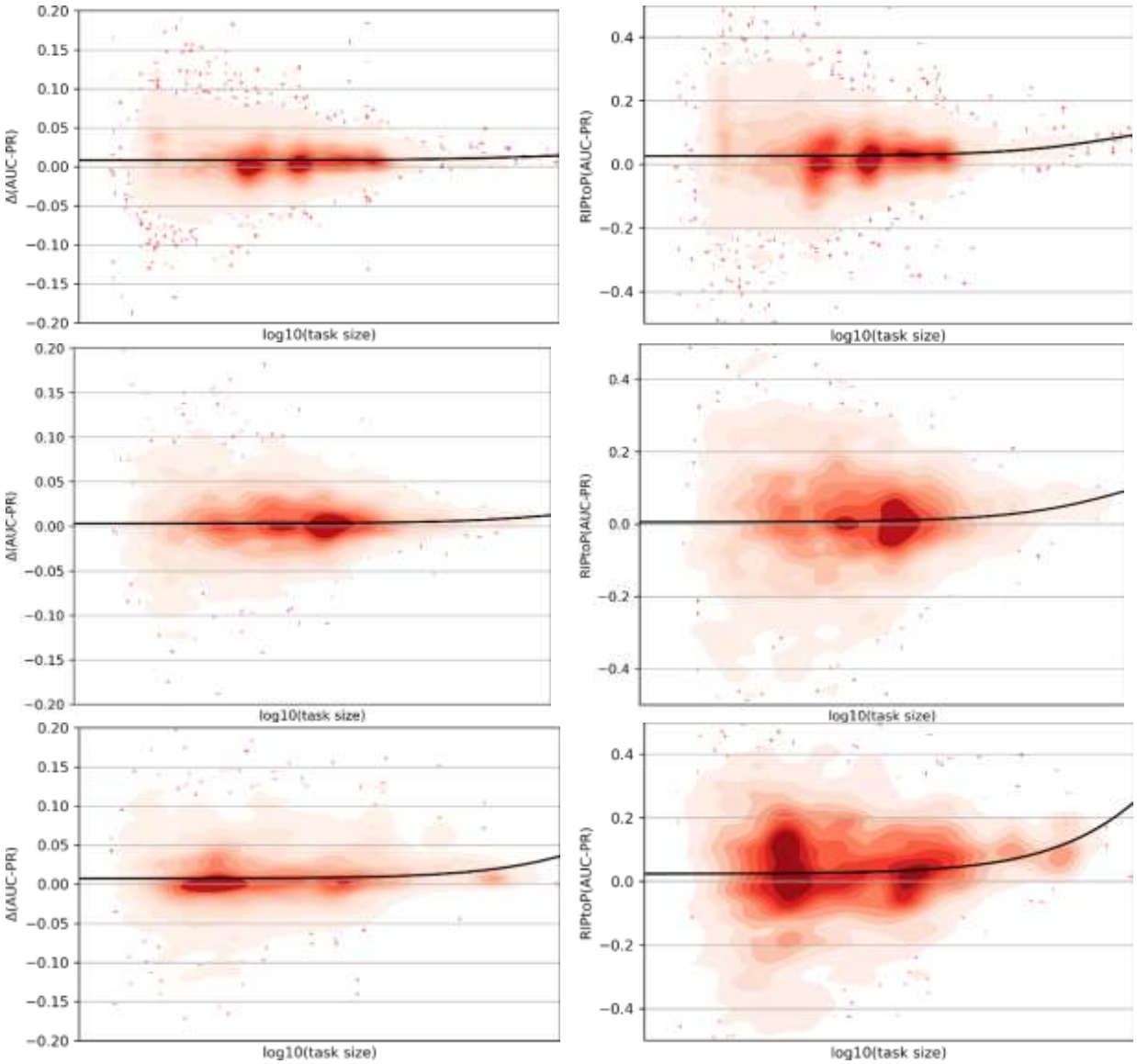


Figure S22. Effect of task size on the absolute (left) and RIPtoP (right) task performance for classification. Each row corresponds to a pharma partner. The black curve is a fitted second-order polynomial to all tasks, ignoring outliers for RIPtoP(AUC-PR). The density of tasks (kernel density estimation) is shown. Outliers for RIPtoP(AUC-PR) between -0.5 and 0.5 were removed from the analysis.

Regression

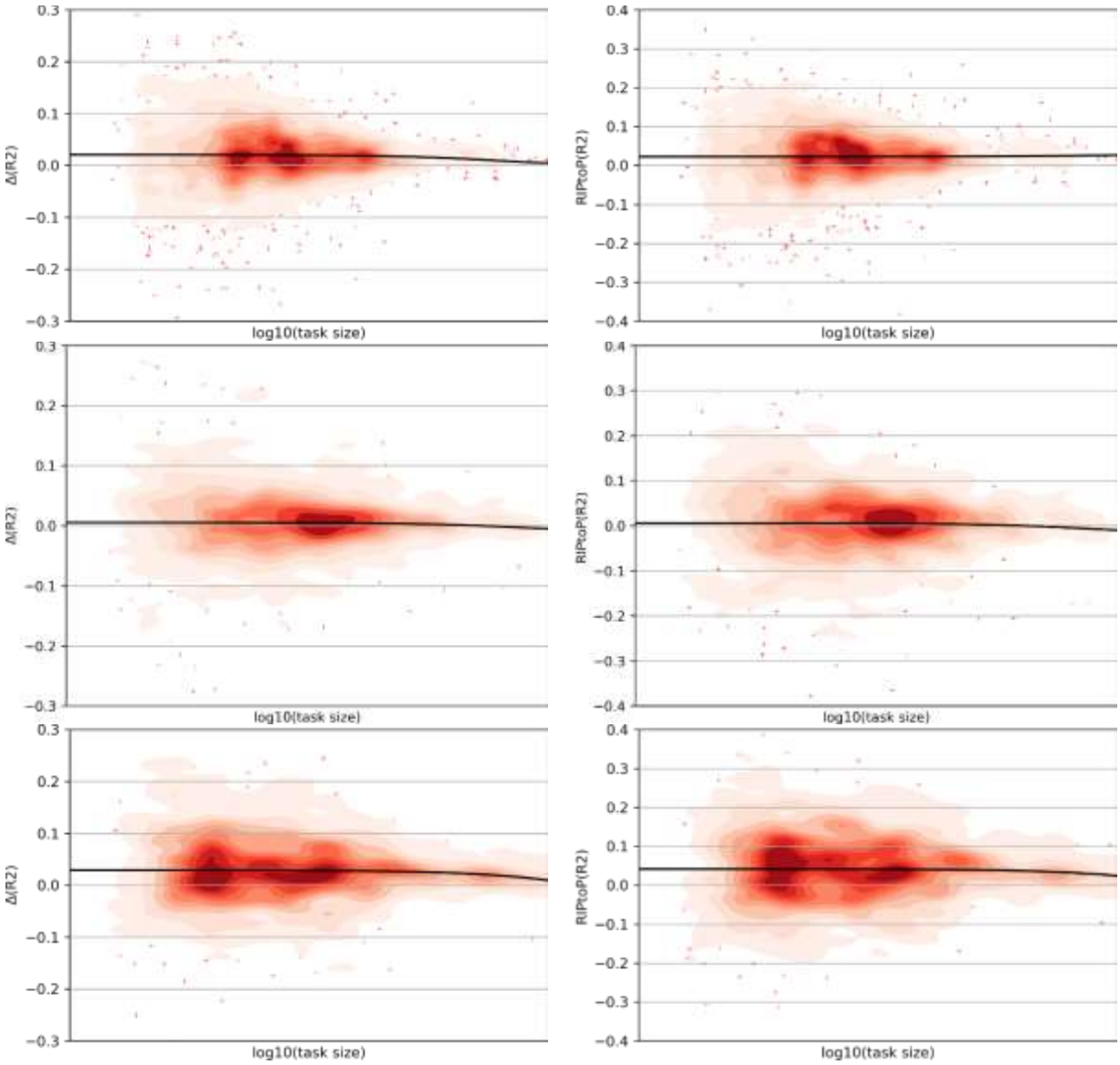


Figure S23. Effect of task size on the absolute (left) and RIPtoP (right) task performance for regression. Each row corresponds to a pharma partner. The black curve is a fitted second-order polynomial to all tasks. The density of tasks (kernel density estimation) is shown.

References

(1) Arany, A.; Simm, J.; Oldenhof, M.; Moreau, Y. SparseChem: Fast and Accurate Machine Learning Model for Small Molecules. *arXiv* **2022**, arXiv ID: 2203.04676.

(2) Gaulton, A.; Hersey, A.; Nowotka, M. L.; Patricia Bento, A.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. https://doi.org/10.1093/nar/gkw1074.

(3) *MELLODDY - Public Data Extraction*. https://github.com/melloddy/public_data_extraction.

(4) Simm, J.; Klambauer, G.; Arany, A.; Steijaert, M.; Wegner, J. K.; Gustin, E.; Chupakhin, V.; Chong, Y. T.; Vialard, J.; Buijnsters, P.; Velter, I.; Vapirev, A.; Singh, S.; Carpenter, A. E.; Wuyts, R.; Hochreiter, S.; Moreau, Y.; Ceulemans, H. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chem. Biol.* **2018**, *25*, 611-618.e3. https://doi.org/10.1016/j.chembiol.2018.01.015.

(5) Hofmarcher, M.; Rumetshofer, E.; Clevert, D. A.; Hochreiter, S.; Klambauer, G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1163–1171. https://doi.org/10.1021/acs.jcim.8b00670.

(6) Heyndrickx, W.; Arany, A.; Simm, J.; Pentina, A.; Sturm, N.; Humbeck, L.; Mervin, L.; Zalewski, A.; Oldenhof, M.; Schmidtke, P.; Friedrich, L.; Loeb, R.; Afanasyeva, A.; Schuffenhauer, A.; Moreau, Y.; Ceulemans, H. Conformal Efficiency as a Metric for Comparative Model Assessment Befitting Federated Learning. *Artif. Intell. Life Sci.* **2023**, *3*, 100070. https://doi.org/10.1016/j.ailsci.2023.100070.

(7) *MELLODDY - Pseudolabel auxiliary data*. https://github.com/melloddy/pseudolabel_auxdata.

(8) *MELLODDY-TUNER*. https://github.com/melloddy/MELLODDY-TUNER.

(9) Landrum, G. *RDKit: Open-Source Cheminformatics Software*. http://www.rdkit.org/.

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

(11) Kruger, F.; Stiefl, N.; Landrum, G. A. RdScaffoldNetwork: The Scaffold Network Implementation in RDKit. *J. Chem. Inf. Model.* **2020**, *60*, 3331–3335. https://doi.org/10.1021/acs.jcim.0c00296.

(12) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree -

Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58. https://doi.org/10.1021/ci600338x.

(13) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing Safety-Related Drug Attrition: The Use of in Vitro Pharmacological Profiling. *Nat. Rev. Drug Discov.* **2012**, *11*, 909–922. https://doi.org/10.1038/nrd3845.

(14) Simm, J.; Humbeck, L.; Zalewski, A.; Sturm, N.; Heyndrickx, W.; Moreau, Y.; Beck, B.; Schuffenhauer, A. Splitting Chemical Structure Data Sets for Federated Privacy-Preserving Machine Learning. *J. Cheminform.* **2021**, *13*, 1–14. https://doi.org/10.1186/s13321-021-00576-2.

(15) 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, P. C.; Friedrich, L.; Formanek, A.; Antal, P.; Rahaman, J.; Zalewski, A.; Heyndrickx, W.; Oluoch, E.; Stößel, M.; Vančo, M.; Endico, D.; Gelus, F.; de Boisfossé, T.; Darbier, A.; Nicollet, A.; Blottière, M.; Telenczuk, M.; Nguyen, V. T.; Martinez, T.; Boillet, C.; Moutet, K.; Picosson, A.; Gasser, A.; Djafar, I.; Simon, A.; Arany, Á.; Simm, J.; Moreau, Y.; Engkvist, O.; Ceulemans, H.; Marini, C.; Galtier, M. Industry-Scale Orchestrated Federated Learning for Drug Discovery. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 15576–15584. https://doi.org/10.1609/aaai.v37i13.26847.

(16) Hu, H.; Salcic, Z.; Sun, L.; Dobbie, G.; Yu, P. S.; Zhang, X. Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.* **2022**, arXiv:2103.07853v4. https://doi.org/10.1145/3523273.

(17) Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* **2021**, *54*, 1–35. https://doi.org/10.1145/3436755.

(18) Acs, G.; Castelluccia, C. I Have a DREAM! (DiffeRentially PrivatE Smart Metering). *13th Inf. Hiding Conf.* **2011**.

(19) Smith, V.; Chiang, C.; Sanjabi, M.; Talwalkar, A. Federated Multi-Task Learning. In *Advances in Neural Information Processing Systems*; 2017.

(20) Agarwal, P.; Sanseau, P.; Cardon, L. R. Novelty in the Target Landscape of the Pharmaceutical Industry. *Nat. Rev. Drug Discov.* **2013**, *12*, 575–576. https://doi.org/10.1038/nrd4089.

(21) Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. https://doi.org/10.1145/3298981.

(22) 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, P. C.; Friedrich, L.; Formanek, A.; Antal, P.; Rahaman, J.; Zalewski, A.; Heyndrickx, W.; Oluoch, E.; Stößel, M.; Vančo, M.; Endico, D.; Gelus, F.; de Boisfossé, T.; Darbier, A.; Nicollet, A.; Blottière, M.; Telenczuk, M.; Nguyen, V. T.; Martinez, T.; Boillet, C.; Moutet, K.; Picosson, A.; Gasser, A.; Djafar, I.; Arany, Á.; Simm, J.; Moreau, Y.; Engkvist, O.; Ceulemans, H.; Marini, C.; Galtier, M. Industry-Scale Orchestrated Federated Learning for Drug Discovery. *arXiv* **2022**, arXiv ID: 2210.08871.

(23) Wenzel, F.; Snoek, J.; Tran, D.; Jenatton, R. Hyperparameter Ensembles for Robustness and Uncertainty Quantification. *Adv. Neural Inf. Process. Syst.* **2020**, *2020-Decem*.

(24) Li, X.; Fourches, D. Inductive Transfer Learning for Molecular Activity Prediction : Next - Gen QSAR Models with MolPMoFiT. *J. Cheminform.* **2020**, 1–15. https://doi.org/10.1186/s13321-020-00430-x.

(25) Imrie, F.; Bradley, A. R.; Van Der Schaar, M.; Deane, C. M. Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. *J. Chem. Inf. Model.* **2018**, *58*, 2319–2330. https://doi.org/10.1021/acs.jcim.8b00350.

(26) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat. Commun.* **2019**, *10*, 1–8. https://doi.org/10.1038/s41467-019-10827-4.

(27) Winter, R.; Montanari, F.; Noé, F.; Clevert, D. A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10*, 1692–1701. https://doi.org/10.1039/c8sc04175j.

(28) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783–790. https://doi.org/10.1021/ci400084k.

(29) Davis, A. M.; Wood, D. J. Quantitative Structure-Activity Relationship Models That Stand the Test of Time. *Mol. Pharm.* **2013**, *10*, 1183–1190. https://doi.org/10.1021/mp300466n.