Collaborative Profile-QSAR: a natural platform for building collaborative models among competing companies

Eric J. Martin* and Xiang-Wei Zhu

Novartis Institute for Biomedical Research, 5300 Chiron Way, Emeryville, California 94608-2916, United States

Abstract:

Massively multitask bioactivity models that transfer learning between thousands of assays work dramatically better than separate models trained on each individual assay. In particular, the applicability domain for a given model can expand from compounds similar to those tested in that specific assay, to those tested across the full profile of contributing assays. If many large companies would share their assay data and train models on the superset, predictions should be better than what each company can do alone. However, a company's compounds, targets and activities are among their most guarded trade secrets. Several strategies have been proposed to share just the individual collaborators' models, without exposing any of the training data. Profile-QSAR (pQSAR), a multitask stacked model which uses profiles of level-1 predictions from single-task models for thousands of assays as compound descriptors for level-2 models, has a uniquely simple and natural fit to collaboration by model sharing.

Broad model sharing has not yet been implemented across multiple large companies, so there are numerous unanswered questions. Novartis was formed from several mergers and acquisitions. In principle, this should allow an internal simulation of model sharing. In practice, lack of metadata about the origins of compounds and assays made this difficult. Nevertheless, we have attempted to simulate this process and propose some findings: multitask pQSAR was always an improvement over single-task models, collaborative multitask modeling did not improve predictions on internal compounds; collaboration did improve predictions for external compounds, but far less than the purely internal multitask modeling for internal compounds; collaborative models for external compounds improve as overlap between compound collections increases; combining profiles from inside and outside the company is never best, with internal predictions best using only the inside profile and external using only the outside profile, but a consensus of models using all three profiles is a good compromise. We anticipate similar results

from other model sharing approaches, most notably the MELLODDY consortium. Indeed, since collaborative pQSAR through model sharing is identical to pQSAR using actual shared data, we believe our conclusions should apply to collaborative modeling by any current method even including the unlikely scenario of directly sharing all the structures and assay data.

Keywords:

pQSAR, model-sharing, stacked model, multitask model, collaborative model, MELLODDY

Introduction

The applicability domains of conventional single-task QSAR models are limited by the chemical space of the compounds tested in that assay. One way to increase the training data is through multitask modeling, where many assays are modeled simultaneously, and learning is transferred among models. This can expand the applicability domain for each assay to approximately cover all the chemical matter in the union of training sets. If companies would pool their compounds and activity data, the applicability domains could be further expanded to cover not just their own compound collections, but those of all collaborators as well. However, collaboration among pharmaceutical companies is rare. Intellectual property and protection of trade secrets are collaboration barriers. Specifically, chemical structures, their biological activities, and each company's targets of interest are carefully guarded. Sharing models without sharing structures or activity data has been proposed for allowing safe collaboration among competitors. The simplest approach to collaboration without sharing chemical structures is consensus modeling for a single assay or target. Under the Medicines for Malaria Venture[1], a consortium of 8 institutions trained naive Bayes binary categorical models on 11 malaria high-throughput screening data sets. A "metamodel" combining 11 Naïve Bayes models performed better than the individual models.[2, 3] The most recent version of that model is now available online.[4] Gedeck *et al.* developed a method to combine partial Bayesian ridge regression models that almost exactly reproduced the results for a single model trained on all the data.[5] Beyond protection of trade secrets, sharing clinical data involves legal, privacy, technical, and data-ownership challenges. Federated semantic segmentation multimodal brain scan models based on a convolutional neural network were similar to models trained by sharing data.[6] The owners do not share their data, but rather

train the shared models locally, and only send model updates to a central server. The server aggregates the updates and sends the new, shared parameters to the data owners for further training or application. However, while consensus models do not share data, they do require revealing the targets. Furthermore, they are single-target models, and the applicability domain is still restricted to chemical matter similar to available training data for that one target.

Federated multitask models can share vastly more data and need not even reveal the targets. The **MELLODDY (M**achine **L**earning **L**edger **O**rchestration for **D**rug **D**iscover**y)** project aims to establish a massively multitask machine learning platform where sensitive data and assay specific models are kept within each partner's firewall.[7]

> The MELLODDY developers intend to create a deep learning model on a centralized server that will travel among these different cloud clusters to train on each company's annotated data, allowing the model to get exposed to a significantly wider range of data than any one company has in house. Once the model has trained for a couple of iterations on an individual company's data, it is sent back to the centralized server to aggregate the contributions before moving on to the next company's data cluster. In essence, the sensitive data remains protected within each individual company's secure infrastructure – it's only the non-sensitive models that are exchanged.[8]

Funded through the Innovative Medicines Initiative (IMI), this 3-year project includes 10 pharmaceutical companies, 7 non-profits and a $20 million USD budget. Ideally, transfer learning between tasks will expand the applicability domain of the multitask models to encompass chemical matter from all collaborator's training sets.

Profile-QSAR (pQSAR) is naturally suited for federated multitask modeling. As illustrated in Figure 1, pQSAR[9-11] is a multitask stacked model. For level-1, conventional single-assay random-forest regression (RFR) models are built on the available pIC_{50} data for each individual assay describing compounds by chemical substructure fingerprints. For level-2, separate PLS models are built for each assay using the profile of predicted pIC_{50}s from all the other level-1 RFR models as compound descriptors. Figure 2 shows that when evaluated on a realistically novel test set, transfer learning across the assays *via* multitask pQSAR dramatically increases the accuracy on compounds very unlike the training set, expanding the applicability domain to roughly cover all the compounds tested across all the assays. A massive pQSAR produced successful models for 72% of nearly 12,000 Novartis continuous-valued assays, with average accuracy on a

realistically novel test set comparable to 4-concentration IC_{50}s, compared to only 8% success for the original single-assay models.

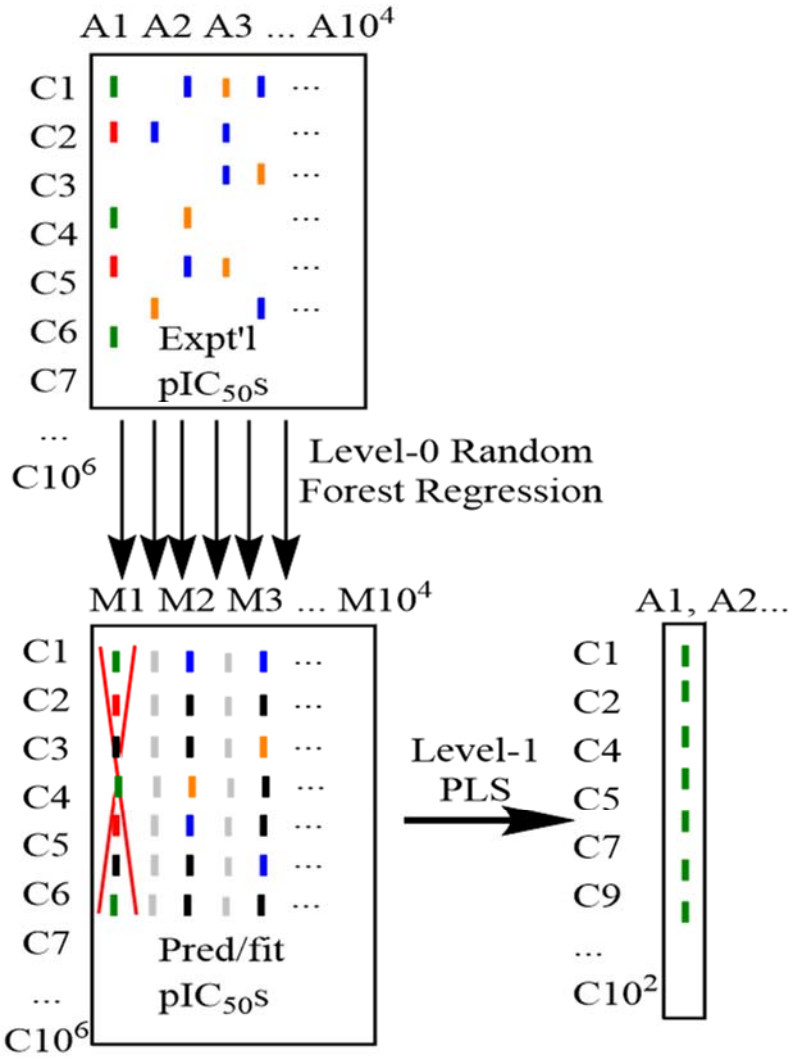


Figure 1. Max2 two-level pQSAR workflow illustrating the training of assay A1. (red = $A1_{test}$, Green = $A1_{train}$, Gold= A^*_{test}, Blue=A^*_{train}, Black = A^*_{NA}). In level-1, single-assay RFR models are trained for all assays using Morgan 2 fingerprints. Level-2 PLS models use the profile of level-1 activity predictions as compound descriptors. The red X indicates that the assay being modeled is not included in the profile. Grey indicates columns excluded from the A1 PLS profile because they do not correlate well enough with A1's training set.

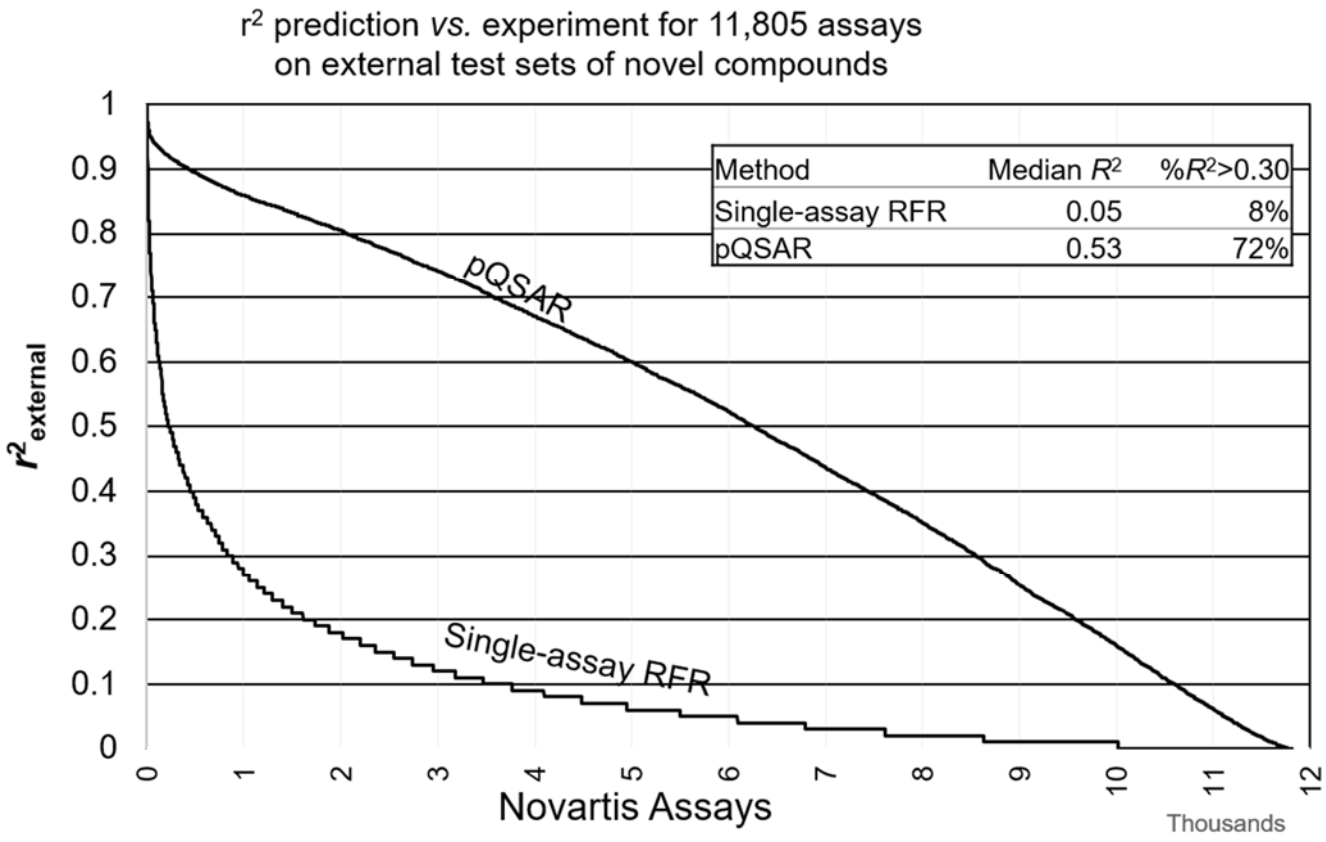


Figure 2. Correlations between prediction and experiment using the "realistic" training/test set split for single-assay random forest regression models and multitask stacked pQSAR models trained using predictions from those very same RFR models as compound descriptors.

For collaborative model sharing (Figure 3), the level-1 single-assay models are trained separately by each collaborator, then shared without revealing the chemical structures, biological activities, targets, or even the type of model. Each single-assay model is just a compiled "black box" that takes SMILES as inputs, and outputs predicted activities. The profile of predictions from the imported level-1 models are used as the compound descriptors to train the level-2 models for each collaborator's own internal assays. While this increases the number of descriptors, the Max2 method reduces each profile to only those single-task level-1 models with some correlation with each assay's training set[1], helping to ameliorate the "curse of dimensionality" as indicated by the greyed-out columns in Figure 1. Note that whereas MELLODDY and other proposed methods are workarounds, using sophisticated procedures different from what one would do if one simply had all the data, collaborative pQSAR uses the exact algorithm the collaborators would use if they had shared the data outright. It just distributes the work among

the partners such that each partner deals with their own data and only gets the final PLS models for their own assays. All the compounds and biological activities used to train all of the collaborators' single-assay level-1 models inform each collaborator's pQSAR models for their own assays identically as if they had actually shared all of the structures and activity data, expanding the accuracy and applicability domains correspondingly.

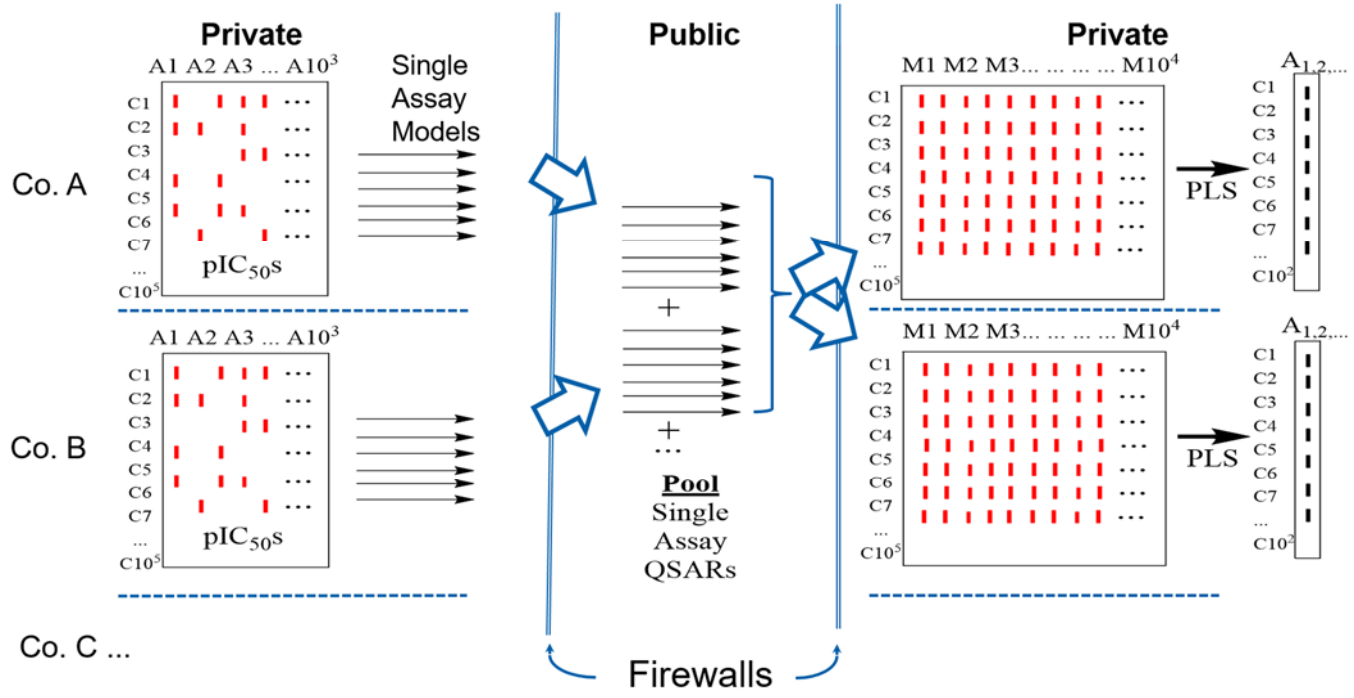


Figure 3. Collaborative pQSAR workflow. Only black-box, single-assay models are exposed outside the collaborator's firewalls.

Broad model sharing has not yet been implemented across multiple large companies, so there are numerous unanswered questions: will adding collaborator data (via models) improve predictions for my compounds on my assays, will it improve predictions for collaborator's compounds on my assays (extended applicability domain), how much must my compound collection overlap with theirs and what is the best way to apply the technology? Novartis was formed from several smaller companies and institutes. In principle, this should allow an internal study of these questions. In practice, lack of compound and assay metadata makes this difficult. Nevertheless, we have attempted to simulate this process and propose some answers to these questions.

Materials and Methods

Detailed descriptions of the procedure for building and evaluating pQSAR models have been previously published. [9-11]

Contributing companies

The mergers and acquisitions that formed Novartis provides an opportunity to simulate collaboration among competitors using the assay data that originated from the precursor companies. The key historical events include Ciba-Geigy (Ciba) and Sandoz merging in 1996 to form Novartis, founding The Genomics Institute of the Novartis Research Foundation (GNF) in 1999, and Novartis acquiring Chiron Corp. in 2006. Even though the compounds and assays have been transferring and evolving across these companies, we did our best to reassemble them into several pseudo-companies through some archaeological study of our database.

Compound assignment

As of September 2019, Novartis AG had nearly 12,000 dose-response assays with at least 50 compounds each and pIC_{50} standard deviation of at least 0.50. Assays included 1.9 million compounds, each of which on average was tested on roughly 10 assays. Unfortunately, the provenance of the assays and compounds were not recorded. Many compounds originating from Ciba, Sandoz or Chiron could be recognized by the formats of their identifiers. Compounds with synonyms matching multiple formats indicate overlap between collections. GNF identifiers were the same as NVS, so compounds registered by GNF employees (unfortunately annotated by sometimes-ambiguous names, not personnel identifiers) were assigned as GNF compounds. The remaining compounds were assigned to NVS. Table 1 shows the overlap between pseudo-companies as identified by synonyms for the same compound. However, the procedures for ensuring one unique entry for every registration of the same compound is not perfect. Besides registration errors, stereochemistry might not be assigned until after registration and canonicalization of tautomers was inconsistently performed. Furthermore, the level-1 RFR models describe compounds by their Morgan 2 fingerprints. Comparing fingerprints can overestimate the overlap, *e.g.* due to bit collisions from folding the fingerprints, and because Morgan 2 fingerprints do not distinguish stereoisomers. Even so, the fingerprint overlaps, shown in Table 2, are generally less than 1%. This is lower than the expected 5 - 10% from recollections of these mergers and anecdotal reports from other company mergers, suggesting problems with our best efforts to classify compound origins retrospectively.

Table 1. The overlap among the five pseudo-companies from compound synonyms.

Co.	NVS	GNF	CHIRON	CIBA	SANDOZ
NVS		27	86	158	202
GNF			3	9	25
CHIRON				44	57
CIBA					245

Table 2. The overlap among the five pseudo-companies by Morgan fingerprints (radius 2, 1024 bits)

Co.	NVS	GNF	CHIRON	CIBA	SANDOZ
NVS		2469	754	1534	1605
GNF			687	330	353
CHIRON				74	96
CIBA					393

Assay assignment

Even worse than compounds, we lacked any annotations for which assays came from which company. Each assay, therefore, was simply assigned to the pseudo-company for which the most compounds were tested. The numbers of assays and principal protein families are shown in Table 3. This presumably overestimates the NVS assays, because an assay originating at a smaller company might have been run on a larger number of NVS compounds after merging. Table 4 shows the distribution of compounds and IC_{50}s within each company's assigned assays. The columns are each company's assays in aggregate. The rows are companies of origin for compounds tested on those assays. Because some compounds were assigned to several companies, the columns do not quite add up to 100%, but the overlap is small so they are close. Compounds originating from the same company as the assay, *i.e.* on the diagonal of Table 4, are "internal compounds". Those assigned to other companies, on the off-diagonals, are "external compounds." *E.g*. thirty percent of compounds tested in NVS assays lack NVS assignments, and are therefore NVS externals. NVS has many more assays than the other four pseudo-companies combined.

Table 3. Summary of assay counts and largest assay families of five pseudo-companies

Co.	#Assays	# 1st Assay family	# 2nd Assay family	#others
NVS	9632	Phenotypic(4904)	Kinase(1728)	3000
GNF	1392	Phenotypic(604)	Kinase(431)	357
CHIRON	305	Kinase(197)	Phenotypic(81)	27

CIBA	280	Phenotypic(116)	GPCR(54)	110
SANDOZ	182	Phenotypic(106)	Receptor (14)	62

Table 4. Distribution of internal and external compounds/IC50s tested in five pseudo-companies

Assays / Cpds	NVS	GNF	Chiron	Ciba	Sandoz
nvs	0.70/0.72	0.30/0.28	0.17/0.11	0.20/0.16	0.23/0.22
gnf	0.13/0.10	0.51/0.57	0.06/0.06	0.03/0.03	0.05/0.06
chiron	0.06/0.05	0.10/0.08	0.72/0.81	0.01/0.01	0.01/0.03
ciba	0.07/0.08	0.06/0.05	0.03/0.01	0.68/0.76	0.09/0.15
sandoz	0.04/0.06	0.03/0.03	0.02/0.01	0.07/0.05	0.61/0.54

Composite pseudo-company "GCCS"

To simulate the collaboration of two big pharma companies of comparable size, we also combined all the compounds from the four small groups (GNF, Chiron, Ciba, and Sandoz) to simulate one larger pseudo-company, "GCCS". The source of the majority of compounds for each of the 12K assays was used to assign assays to the two pseudo-companies. NVS retained 8297 assays with 1.5M compounds and 14M activities. GCCS had 3494 assays with 0.77M compounds and 4.7M activities. Varying percentages of compound overlap between NVS and GCCS were created by several methods: moving varying numbers of assays from NVS to CCGS and adding GCCS labels to 90% of the moved compounds, moving 1000 assays from NVS to GCCS and adding GCCS labels to different fractions of the moved compounds, and adding GCCS labels to different fractions while deleting the same number of compounds from the original GCCS assays (see below).

Training and test sets

As illustrated in Figure 4, for each company, for each assay inside that company, compounds assigned to that company form its internal set and those assigned only to other companies form its external test set. Predictions on the external test set represent the task of predicting compounds outside the company archive. Internal compounds were split into a training and an internal test sets by the "realistic" algorithm, assigning the 75% of compounds from the largest clusters for training, and holding out the remaining 25% of singletons and small clusters for testing. This very challenging test set is called "realistic" because it was shown to mimic the

extreme novelty of compounds that project teams chose to test in real virtual screens.[10] The value of collaboration was tested by comparing, for each assay, the prediction quality of pQSAR models built on three profiles (sets of level-1 single-assay RFR pIC_{50} predictions used as PLS compound descriptors): 1) using only RFR models from inside the company, 2) using only outside RFR models from the partners, and 3) combining all RFR models both from inside and outside. To minimize confusion, "inside" *vs.* "outside" refers to assays or level-1 RFR models, while "internal" *vs.* "external" refers to compounds. The workflow was as follows: 1) ~12,000 level-1 random forest regression (RFR) models were trained using 100% of the internal compounds for each assay; 2) ~12,000 level-2 PLS models were then built on the 75% training sets of each assay, using as descriptors the predicted activities from the RFR models in each of the 3 profiles *except* the current assay being modeled (which avoids test-set leakage) and 3) performance was tested on that assay's realistic internal test set to assess internal performance, and on that assay's external test set to assess external performance. Assays with fewer than 3 external compounds were not used for evaluating either external or internal predictions, but were used in the profiles. Bear in mind that typically, internal compounds were tested in several inside assays (and potentially in some outside assays in the cases of collection overlap), so their predictions with inside or combined profiles are imputations, *i.e.* their "predictions" by some of the level-1 single-assay RFR models are actually fits. This is typical of real virtual screens of a corporate archive. Compounds in the external test sets, by definition, were not used in training the inside profiles, although many were used in training the outside profiles. Hence, they could only be imputations when using the outside and combined profiles. The performance on the internal or external test sets using the inside profile alone shows the performance of non-collaborative pQSAR models in virtual screening. Better performance with outside or combined profiles shows a benefit to collaboration. Table 5 shows the number of assays, internal and external compounds in each pseudo-company (rows) and average frequency they were tested in inside and outside assays. For example, the first row starts with the number of NVS assays and total number of compounds tested in them. It shows how many compounds tested in NVS assays were internal and external, and the average frequency of testing NVS internal compounds on assays inside NVS and NVS external compounds on assays outside NVS. larger numbers suggest a larger fraction of imputations. Note that NVS external compounds were rarely tested outside NVS, but for the smaller companies the opposite is true.

Table 5. The counts and test frequencies of compounds and assays internal and external to each pseudo-company.

Co.	#Assay	#compds	Int-Cnt	TFI	Ext-Cnt	TFO
NVS	10708	1684121	1179329	11.44	504792	3.27
GNF	1530	411309	207841	5.21	203468	27.15
Chiron	329	118279	84599	9.59	33680	60.22
Ciba	324	84874	58121	2.92	26753	48.08
Sandoz	200	43270	26549	2.83	16721	49.83

Int-Cnt: Count of internal compounds
TFI: Test frequency of internal compounds on inside assays
Ext-Cnt: Count of external compounds
TFO: Test frequency of external compounds on outside assays

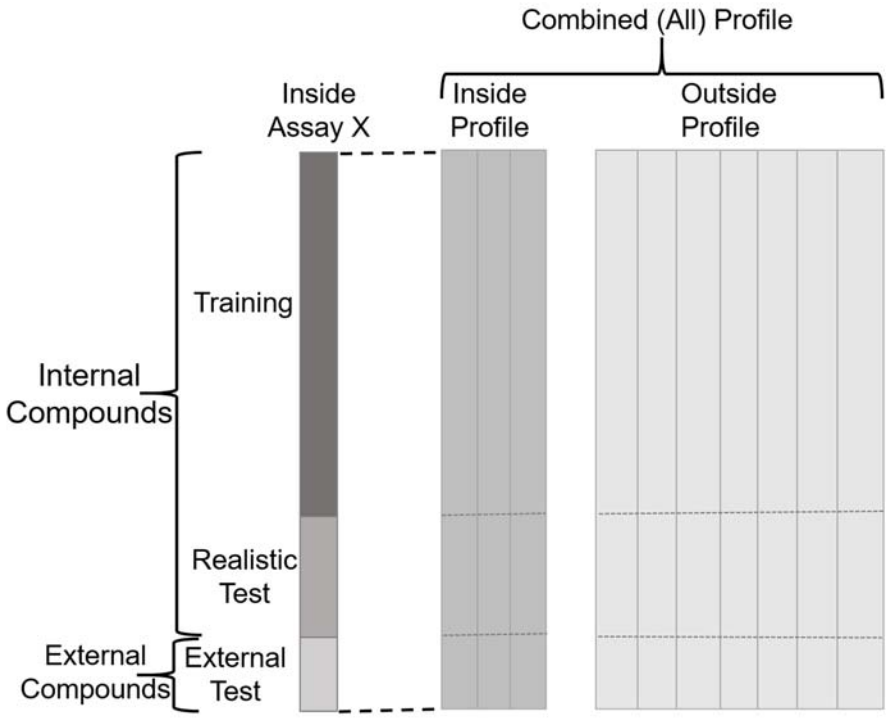
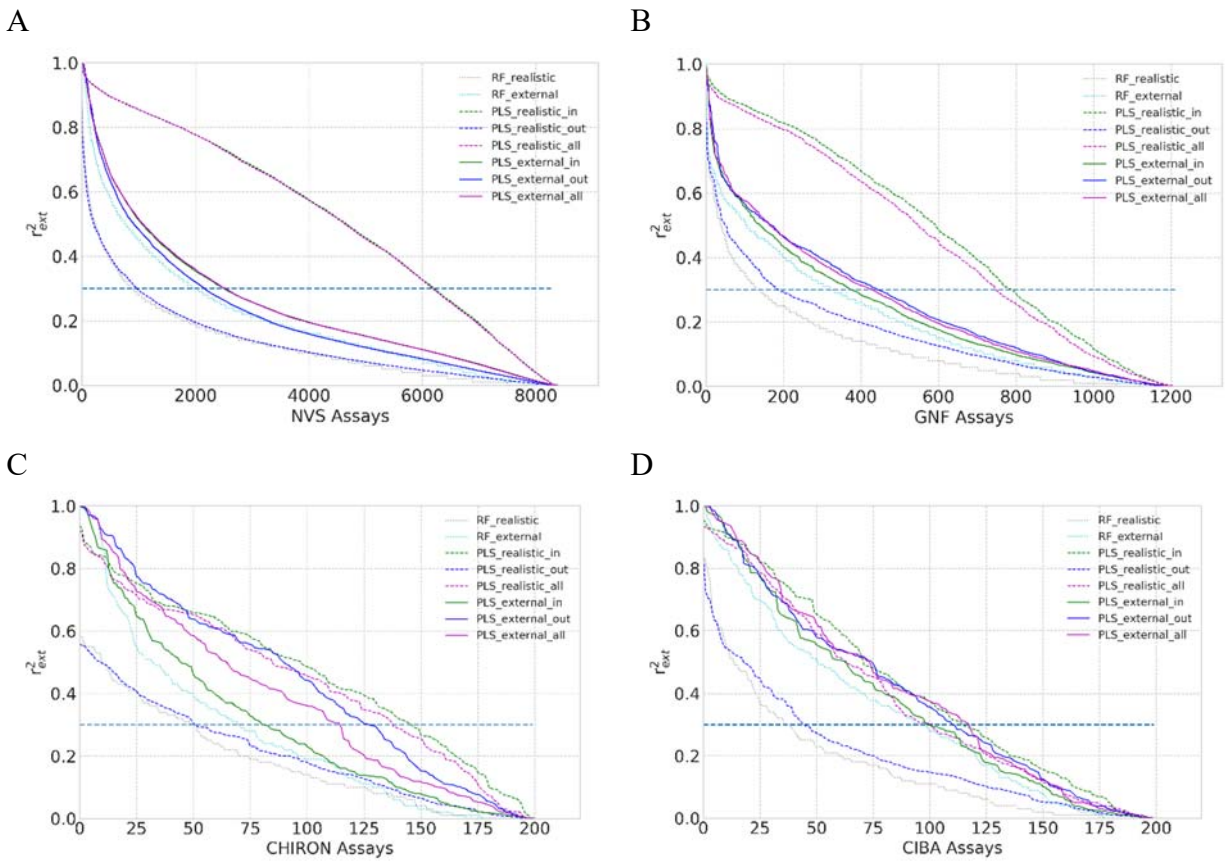


Figure 4. Illustration of dataset splitting and collaborative PLS model training. For each company's inside assays, three pQSAR models are trained by using three, level-1 RFR profiles for the level-2 PLS compound descriptors: using only that company's RFR models (in), using only the collaborators' RFR models (out), and using all RFR models.

Results

Case 1. Collaboration among five pseudo-companies

In Case 1, NVS, GNF, Chiron, Ciba, and Sandoz all contribute their level-1, single-assay RFR models to a common pool. Each partner has access to all the RFR models. Rank-order plots in Figure 5 compare predictions on realistic internal and external test sets using pQSAR models with inside, outside, and all profiles, as well as the RFR single-task models. To simplify presentation, the pQSAR curves are also characterized in Table 6 and Figure 6 by median r^2 between prediction and experiment, and percentage of assays with r^2 >0.3.



E

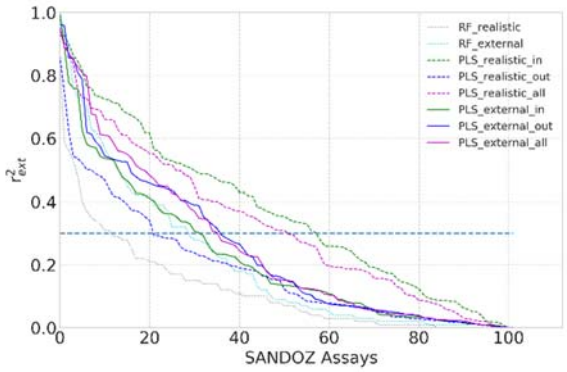


Figure 5. Rank-order plots of r^2 for prediction vs. experiment on 2 test sets (external and realistic internal) for single-assay RFR models, and for pQSAR models using 3 level-1 RFR profiles: inside RFR models only (in), outside RFR models only (out) and combining inside and outside RFR models (all). The dashed horizontal line indicates our success threshold of $r^2 = 0.30$.

Single-assay RFR r^2 is always higher on the external test set than on the realistic internal test set. This illustrates how challenging the realistic test set of singletons and tiny clusters is. *I.e.* the external test sets are more like the inside training sets than are the realistic internal test sets. This is quantified in frequency tables in Supplemental **Error! Reference source not found.**.

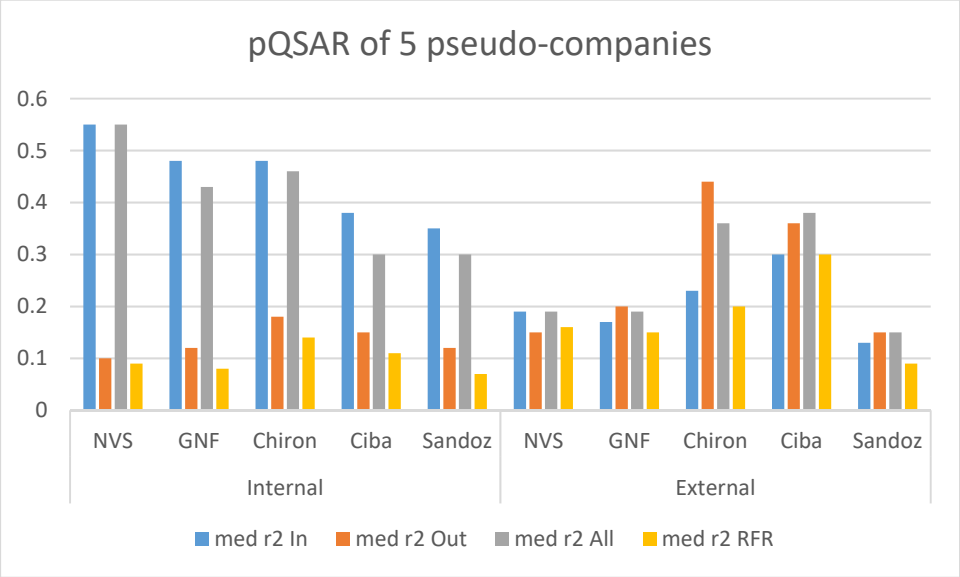
The order of pQSAR profiles for the realistic internal test sets is the same in all cases: inside >= combined >> outside (dashed lines in Figure 5, left side in Figure 6). The improvement over the corresponding RFR models on internal realistic test sets from pQSAR models with inside or combined profiles is dramatic. As mentioned above, many compounds or their near analogs in the realistic internal test set for each assay have been tested in other assays in the inside profile. These imputations or near imputations benefit most from transfer learning. However, pQSAR is also better than RFR on the internal test set even using just the outside profile, which includes very few imputations.

Looking in more detail at Table 6, NVS performance of level-1 RFR models on the internal realistic test set is very weak, with median $r^2 = 0.09$ and just 11.1% of assays achieving $r^2 > 0.3$, our general rule for a successful virtual screen. Profile-QSAR models using just the RFR profile from the relatively few outside partner models is nearly as poor. pQSAR performance on the internal realistic test set using just inside RFR models is outstanding, with median $r^2 = 0.55$ and fully 74% of assays achieving $r^2 > 0.3$. Adding RFR models from its much smaller partners into the profile (*i.e.* All) had no effect. The pQSAR predictions for internal compounds from the 4 smaller pseudo-companies are also impressive. Even in these cases, using just the small inside

profiles is better than using the huge "all" profile that includes much larger NVS. This might be because the simplistic Max2 variable reduction step does not fully compensate for the curse of dimensionality due to adding so many NVS models to the profile. Thus, collaboration never improved predictions on internal compounds, and adding more collaborator models actually made it worse.

Table 6. Median r^2 and % r^2 > 0.30 for pQSAR models built using 3 RFR profiles: inside only (in), outside only (out) and combined inside and outside (all), on the internal and external test sets for each of the five pseudo-companies, which have almost no overlap among them, as well as single-task RFR models.

Co.	Test set	median r^2				$\%r^2$ >0.30			
		RFR	In	Out	All	RFR	In	Out	All
NVS	Realistic Internal	0.09	0.55	0.1	0.55	11.1	74.0	11.6	73.8
GNF		0.08	0.48	0.12	0.43	11.8	65.1	15.4	61.6
Chiron		0.14	0.48	0.18	0.46	25.5	72.6	25.4	68.7
Ciba		0.11	0.38	0.15	0.3	19.5	57.5	22.5	51.0
Sandoz		0.07	0.35	0.12	0.3	11.8	56.9	20.6	50.0
NVS	External	0.16	0.19	0.15	0.19	24.8	30.0	25.8	30.1
GNF		0.15	0.17	0.2	0.19	27.6	30.8	36.6	35.6
Chiron		0.20	0.23	0.44	0.36	35.5	40.8	63.7	57.2
Ciba		0.3	0.3	0.36	0.38	49.5	49.5	55.5	59.0
Sandoz		0.09	0.13	0.15	0.15	28.4	31.4	35.3	35.3



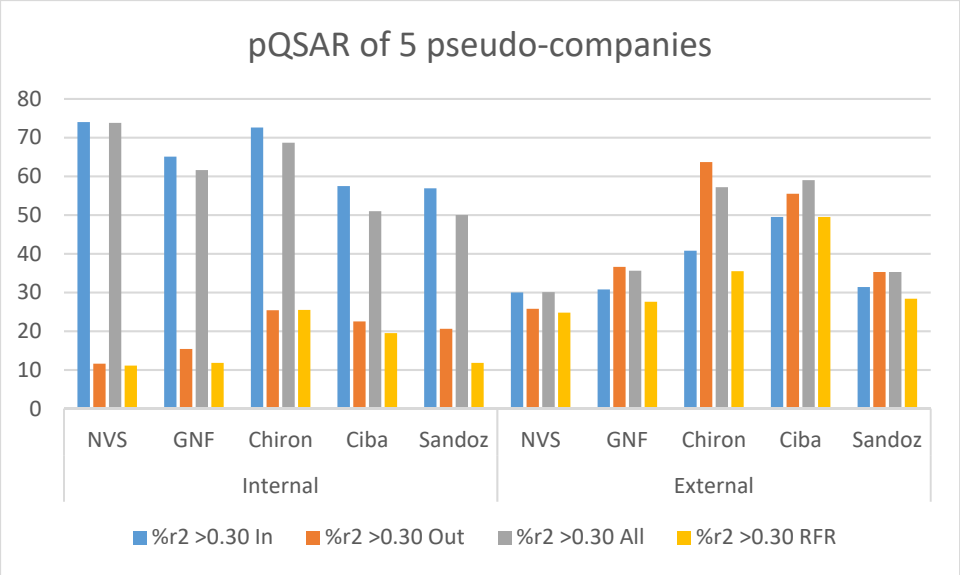


Figure 6. Median r^2 and % r^2 > 0.30 for pQSAR on the internal (realistic split) and external test sets for each of the 5 pseudo-companies which have almost no overlap among them. pQSAR models are compared using 3 profiles: inside RFR models only (in), outside RFR models only (out) and combined inside and outside models (all), as well as single-assay random forest regression models.

For the external test sets, the improvement of pQSAR over single-assay models is real, but much more modest (solid lines in Figure 5, Table 6, right side of Figure 6). For NVS the large inside and combined profiles are comparable, indicating no benefit from collaboration with smaller companies, and are better than much smaller outside alone. For the 4 smaller pseudo-companies on external compounds, the relatively small inside profiles are never best compared to the much larger combined or outside profiles which include NVS. Overall, the benefit of pQSAR is less on external compound predictions than on internal, but collaboration is clearly helpful for the smaller companies, although not for much larger Novartis.

Case 2. Subsample NVS

Concerned that the large NVS profile was dominating the analysis, we randomly sampled just 540 of the NVS assays, the average number of assays in the four smaller partners. The subsampling was replicated 3 times with virtually identical results as shown in **Error! Reference source not found.**, **Error! Reference source not found.**, Figure S1 and the rank-order plots in **Error! Reference source not found.**. One replicate from Figure S1 is shown as Figure 7. In all cases in this study, percentage with r^2>0.3 mirrors mean r^2, so bar graphs for only the latter will be shown from here on.

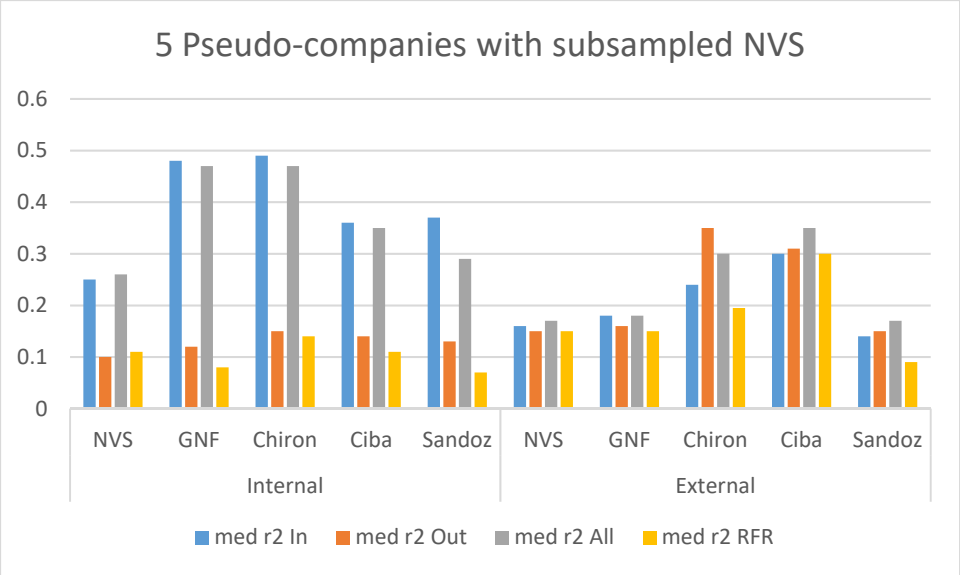


Figure 7. pQSAR median correlations of prediction with experiment on the "realistic" test sets of compounds internal and external to each of the 5 pseudo-companies using a subsampling of just 540 assays for NVS. pQSAR models are compared using 3 profiles: inside, outside and combined (all), as well as single-assay random forest regression models.

The results in Figure 7 are generally similar to Figure 6. For the internal test sets, pQSAR is again far better than single-assay models with inside > combined >> outside, except for NVS where combined is now slightly better than inside. However, the benefit of pQSAR for NVS is much decreased. This might indicate a very high diversity in such a small sampling of the many NVS assays and compounds, with therefore very little correlation, and which is not spanned even the combined profile from all small companies.

For the external test sets, the improvement due to pQSAR again was real but now more modest, indicating that the larger set of helper assays from full NVS had outweighed the noise from so many PLS descriptors. The order of profiles for each pseudo-company was the same as with the large NVS scenario, except that the diminished outside profile now is worst rather than best for GNF, now the largest partner. All worked best except for Chiron, for which using the outside profile alone was best.

Overall, for companies of similar size, even with very little compound overlap, collaboration did generally benefit predictions on external compounds, especially for Chiron.

Case 3. Collaboration between two large pseudo-companies with varying overlap

As mentioned above, lack of annotation frustrated efforts to identify overlapping compounds between pseudo-companies. For example, NVS contains the compounds that were not otherwise assigned, and so should have no overlap, although minor problems in the database did cause a small amount. Yet overlap is key to knowledge transfer in multitask modeling. Anecdotal evidence suggests that when companies merge they find 5-10% of their compounds are in common.

Realizing that we could not do a historically accurate analysis, we tried to create a more controlled comparisons between two large pseudo-companies by artificially engineering overlap. To create artificial compound overlap while reducing company size bias, the GNF, Ciba, Chiron, and Sandoz assays were merged as one large "GCCS" pseudo-company with ~3500 assays. Compound overlap was then systematically introduced by moving assays from NVS to GCCS. We moved the 0, 150, 250 or 500 of the ~8300 NVS assays with the least overlap between the companies. GCCS labels were then added to ninety percent of the moved compounds before model training. Thus, these compounds now had both GCCS and NVS labels. The remaining 10% retained just the NVS label to serve as GCCS external compounds. Finally, the NVS label was deleted from any GCCS external compounds in the 3500 original GCCS assays that were among the newly relabelled GCCS compounds, so they moved to the internal test sets in all assays, thus keeping the external test sets free of internal compounds. This produced four NVS-GCCS overlap scenarios with approximately 0, 5, 10, and 15% overlap.

Figure 8, **Error! Reference source not found.** and **Error! Reference source not found.** show that again multitask pQSAR models were always better than single-task RFR. Improvement over single-task models for the internal test set, with inside or combined profiles, is again dramatic. Increasing overlap by moving assays had little impact on predictions for the internal test sets, even slightly hurting GCCS. As before, collaboration never improved internal predictions.

External predictions also always benefited Profile-QSAR. Using only the outside profile is always best, and performance clearly improves with increasing overlap, especially for GCCS. Yes, even with 15% overlap it was far less than for internal. Interestingly, for GCCS external predictions, the inside profile also gives better performance with increasing overlap, although not for NVS. This could be due to the increasing proportion of "external-like" relabeled compounds

among the GCCS internals when more assays are moved, or simply due to the increasing numbers of assays in the GCCS inside profile.

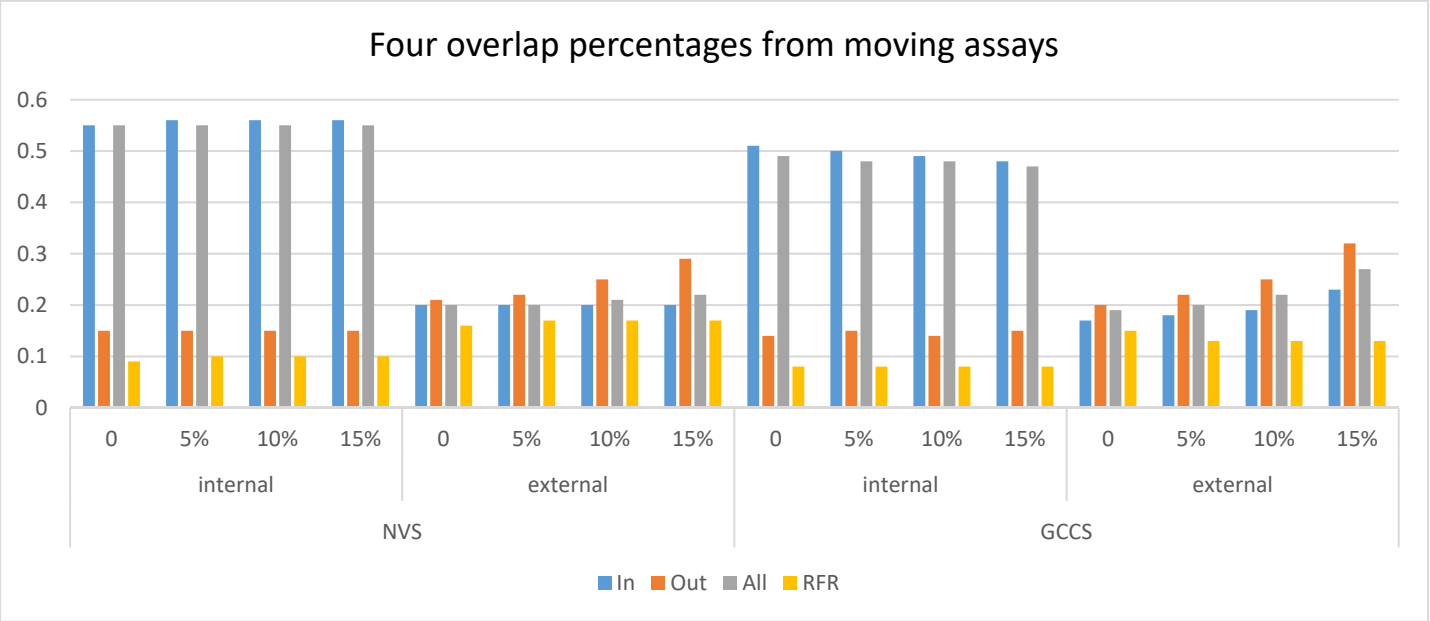


Figure 8. Correlations of prediction vs. experiment on the internal and external test sets of the NVS and GCCS pseudo-companies for single-assay RFR models and for pQSAR models using 3 profiles: inside, outside and all. 0, 5, 10 and 15% overlap was engineered by moving from NVS to GCCS 0, 150, 250 or 500 assays respectively.

Case 4. Collaboration between two pseudo-companies with identical external test sets.

As more assays were moved from NVS to GCCS, the GCCS external compound set increased. To alleviate differences in GCCS external compound sets between overlap scenarios, the analysis from the calculation was limited to the GCCS external compounds common among all four overlap scenarios for each assay. There were 2564 assays common to all 4 GCCS overlap scenarios that had at least 3 common external compounds. The median r^2 for just these common external GCCS compounds is shown in Figure 9 and **Error! Reference source not found.**. The overall correlations with experiment are worse for this GCCS subset, possibly due to some very small test sets, but they still improve with overlap. The inside profile continues to show better performance with increasing overlap for GCCS external predictions, suggesting differing test sets did not account for that trend.

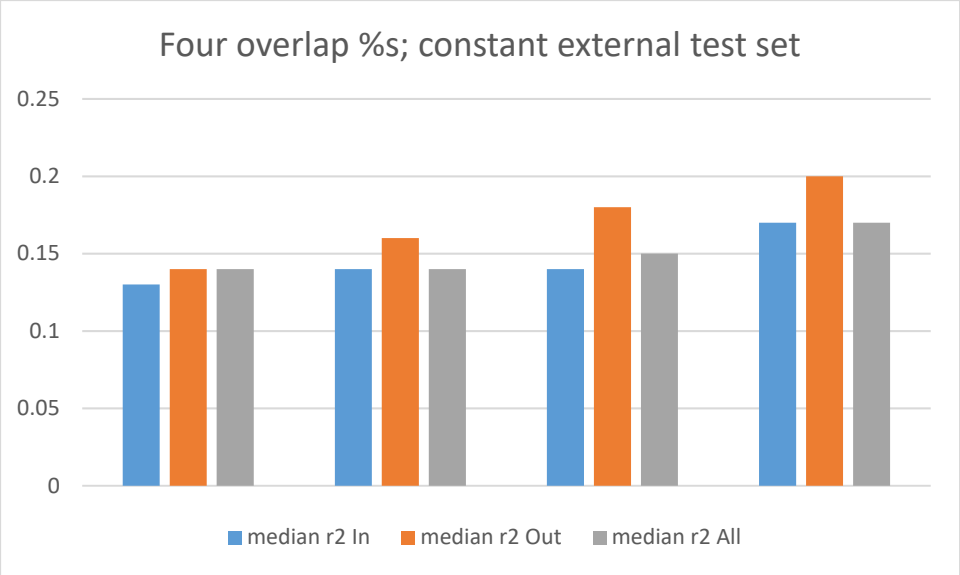


Figure 9. Correlations of prediction vs. experiment from Figure 8Figure 7 analyzing only the GCCS external test compounds common to all 4 overlap scenarios.

Because overlap was made by moving NVS assays to GCCS, almost all of the overlap is in those new 150, 250 and 500 assays. Figure 10 and **Error! Reference source not found.** show the superior model performance for those moved assays compared to the original GCCS assays. The average performance of moved assays does not improve with increasing overlap, whereas the original assays do improve slightly. Thus, the overall improvement is due both to an improvement in the original assays with increasing overlap, and an increasing proportion of good models from moved assays.

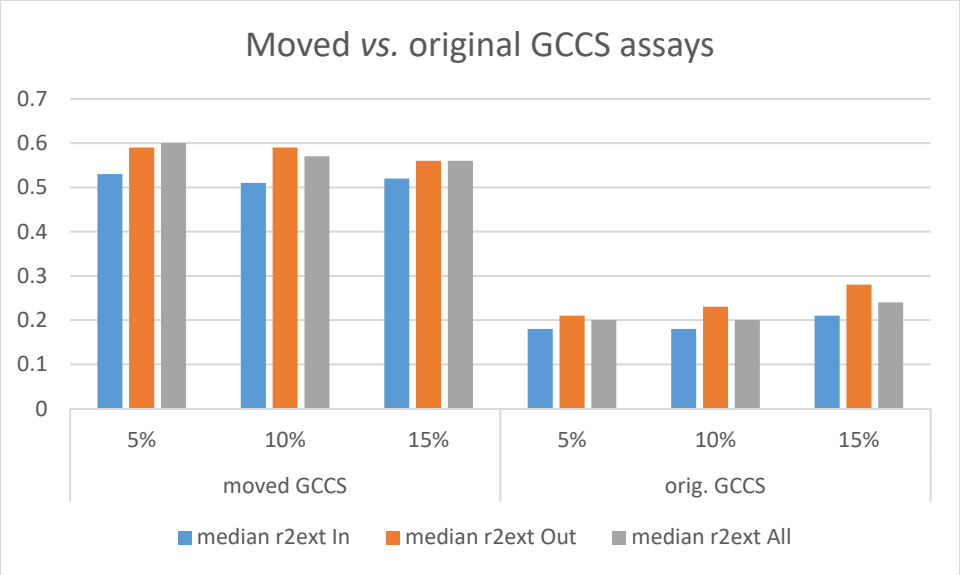


Figure 10. Same as Figure 9 but divided between the moved assays and the original assays.

Case 5. Collaboration between two pseudo-companies with a fixed number of assays.

Since increasingly more assays had been moved to adjust the overlap, the increasingly large GCCS profiles could confound the conclusions about overlap. To construct collaboration scenarios in which the 2 companies have constant sets of assays but different overlap, we first moved 1000 assays from NVS to GCCS. The ratio of adding GCCS labels to NVS compounds was then varied to 20, 50, and 70%, constructing new NVS and GCCS companies with 4, 11, and 15% overlap. Assays with too few external compounds after moving 70% were eliminated, leaving NVS and GCCS with fixed sets of 6573 and 3565 assays respectively.

Figure 11 and **Error! Reference source not found.** show that with constant assays, GCCS external predictions still improve as compound overlap increases. Again, collaboration does not improve internal predictions. Outside-only profiles again substantially beat combined for external predictions. GCCS external predictions improve with overlap for all 3 profile types.

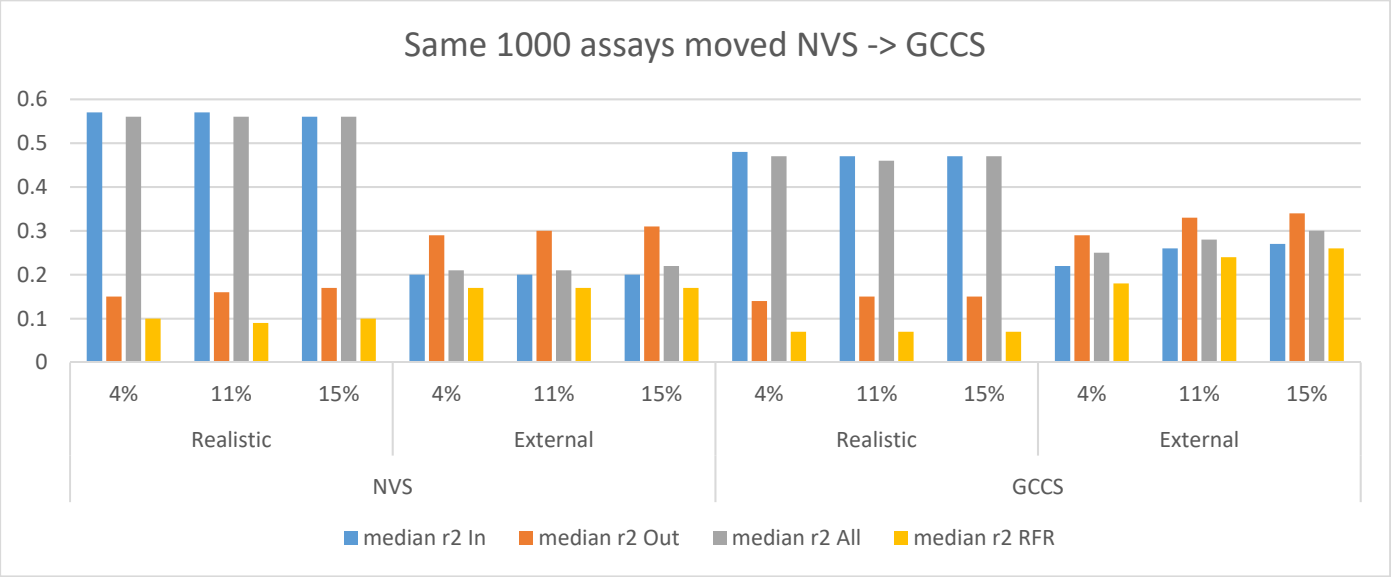


Figure 11. Correlations of prediction vs. experiment on the internal and external test sets of NVS and GCCS for single-assay RFR models and pQSAR models using 3 profiles: inside, outside and all. 4, 11 and 15% overlap was engineered by moving 1000 assays from NVS to GCCS, then reassigning 20, 50 or 70% of the formerly NVS compounds as GCCS internal.

Figure 12, **Error! Reference source not found.** and **Error! Reference source not found.**, analogous to Figure 9, **Error! Reference source not found.** and Figure S3, analyze only the external GCCS compounds common to all 4 overlap scenarios, now with constant RFR profiles. Correlation with experiment is less for this compound subset, but still improves with overlap. Outside still beats combined.

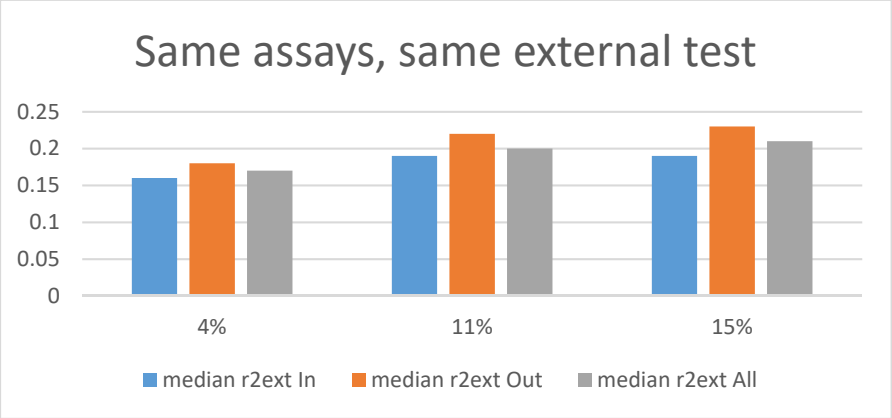


Figure 12. Same as Figure 9, but overlap created by reassigning different numbers of compounds from NVS to GCCS from the same set of 1,000 moved assays.

Figure 13 and **Error! Reference source not found.** show that most of the improvement is in the 1000 models that were moved from NVS to GCCS, where the overlap is concentrated. In this case, the set of moved models is constant, and the overall improvement is due to average improvement with increasing overlap of the fixed set of 1000 moved models. Unlike the original GCCS assays, outside is not better than combined for the moved assays. It does make sense that as overlap increases, both inside and outside RFR models would increasingly help, and these assays have very high overlap.

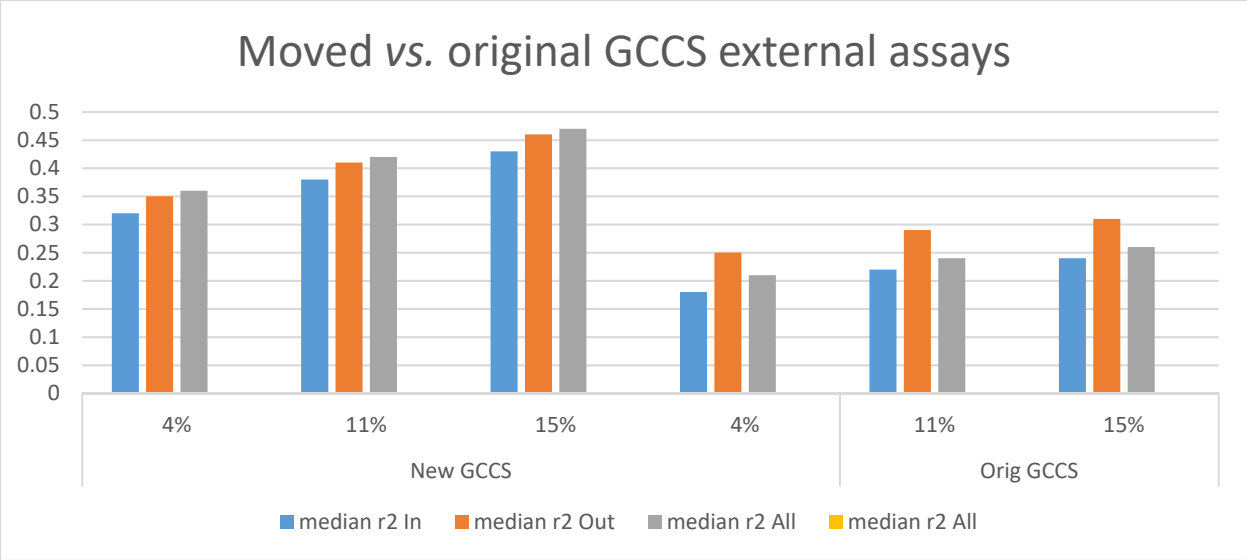


Figure 13. The pQSAR models from Figure 11 for the 1000 moved assays, where most overlap resides, show the best performance, and it increases with overlap.

However, although the moved assays are now fixed, as ever more NVS compounds are reassigned from GCCS external to GCCS internal, the GCCS training data sets increase, which still confounds the interpretation of overlap.

Case 6. Collaboration between two pseudo-companies fixing both the profile assays and the number of internal compounds in GCCS.

To maintain a constant GCCS internal compound set size, when GCCS labels were added to compounds from the 1000 moved assays to increase overlap, the same number of non-overlapping GCCS compounds were deleted at random from the original GCCS internal sets. After some trial and error, we managed to construct new NVS and GCCS companies with 5, 13, and 23% overlap. Thus, both the profile assay sets and the number (although not the identity) of GCCS internal compounds was constant.

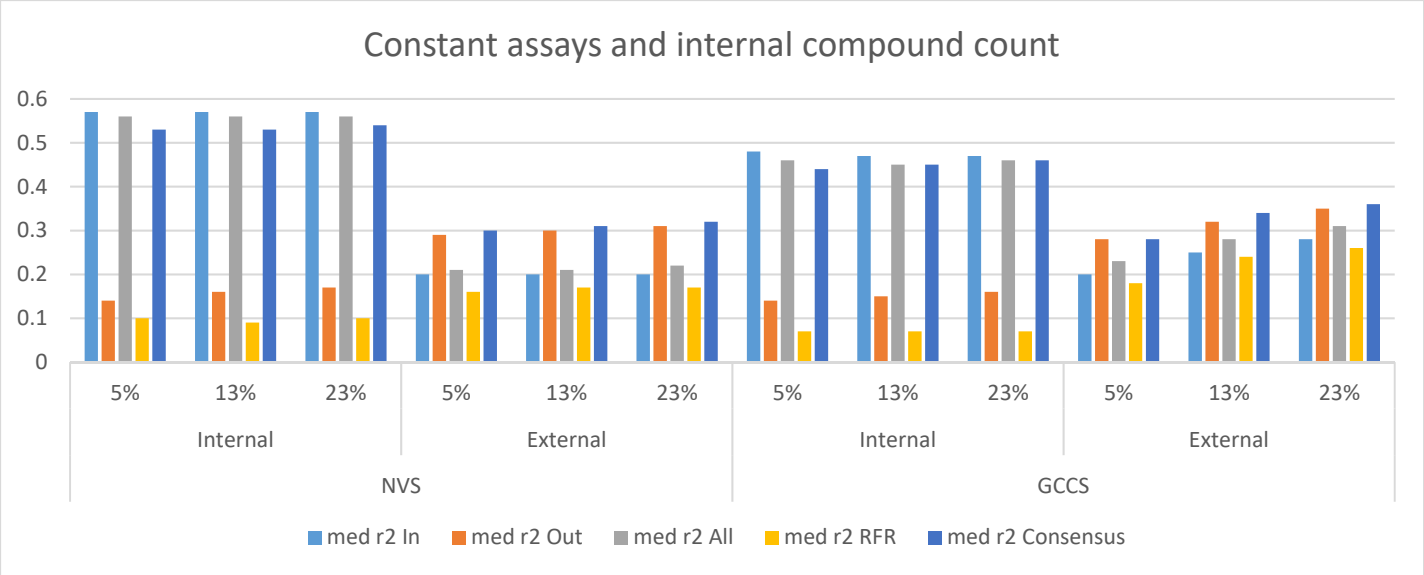


Figure 14. Analogous to Figure 11, but internal compounds are removed from the original GCCS assays to keep the number of training compounds constant.

Figure 14, Figure 15, and Figure 16 (with corresponding Tables S10-S12 and Figure S5) are analogous to Figure 11, 12, and 13. While the overlap percentages are slightly different, the trends are qualitatively the same. Thus, in all 4 methods of comparing the NVS and GCCS pseudo-companies, the benefit of collaboration on external predictions increases with increasing overlap in their compound collections, and this cannot be otherwise explained by changes in assays, training sets sizes, or external test sets.

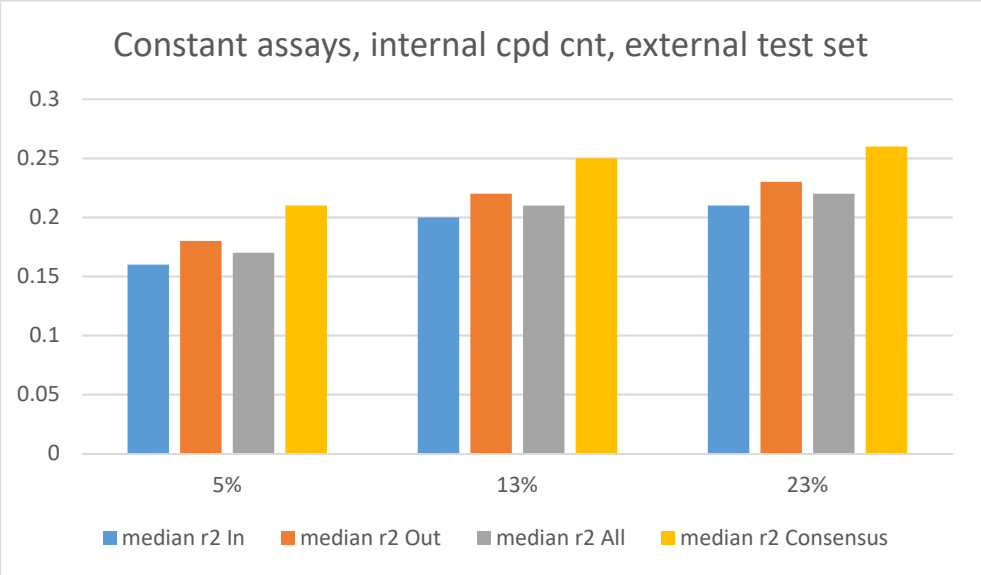


Figure 15. Analogous to Figure 12, but internal compounds are removed from the original GCCS assays to keep the number of training compounds constant.

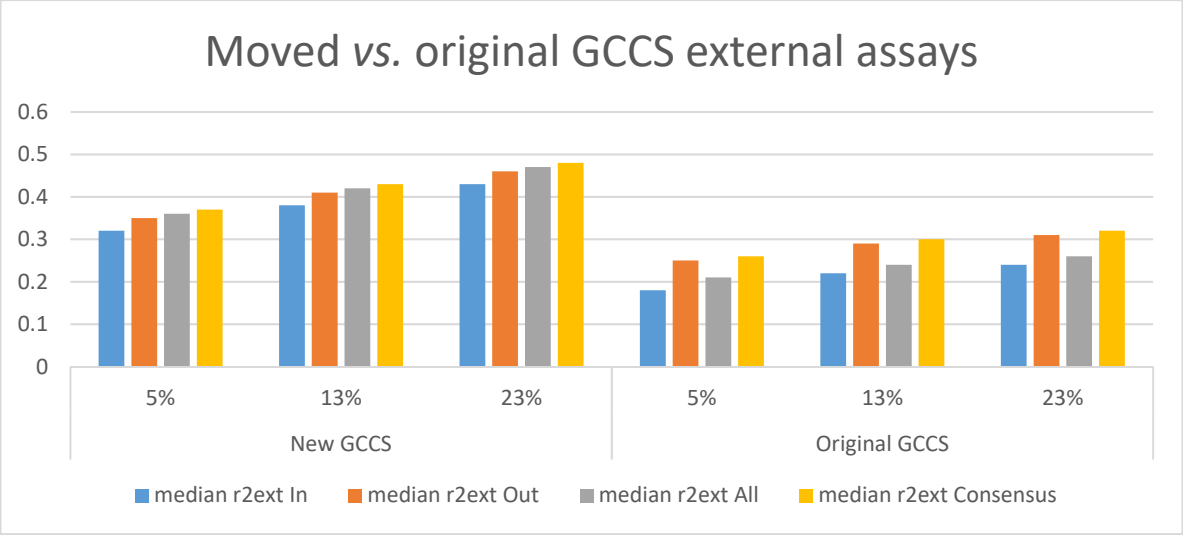


Figure 16. Analogous to Figure 13. The pQSAR models from Figure 11but internal compounds are removed from the original GCCS assays to keep the number of training compounds constant.

Discussion

Comparing Profiles

The pQSAR models built with the outside profile alone were consistently better than the combined profile for predictions on external compounds, except for some moved assays that had exceptionally high overlap. Beyond the problem of adding additional descriptors, other possible explanations are bias in PLS coefficient training and biased variable reduction in the PLS step. By design, the external compounds had no overlap with the training sets for the inside RFR models, but often do overlap with outside RFR training sets. Single-assay RFR models have narrow applicability domains. External compounds will frequently be outside the reliable chemical space of the RFR models from the inside profile even if real experimental measurements for those inside helper assays would correlate well with that assay, and therefore improve the PLS model. External compounds are more likely be within the applicability domains of outside RFR models, so their relevant correlations will more likely help the PLS models. Thus, even in pQSAR models for inside assays, the inside RFR models might add more noise, and the outside RFR models more signal, to the pQSAR predictions on external compounds.

Variable selection during model training reinforces this bias. As Table 7 shows, for pQSAR models using all RFR models in the profile (threshold of r^2>0), the ratio of inside/outside profile models is, of course, the same as the ratio of assays, 65/35 for NVS and thus 35/65 for GCCS.

Max 2 pQSAR uses variable retention thresholds of either $r^2>0.05$ or $r^2>0.20$.[2] For PLS models built with a RFR retention cutoff of $r^2>0.05$, the average ratio became 69/31 for NVS PLS models and 37/63 for GCCS PLS models. At the cutoff of $r^2>0.20$, the ratios were 80/20 for NVS and 45/55 for GCCS. Thus, variable reduction on the combined profiles preferentially eliminates outside models, for which more external compounds would be within the applicability domain. However, Figure 17, **Error! Reference source not found.** and **Error! Reference source not found.** show results for pQSAR models trained using the full profile in the PLS step without Max2 profile reduction for the 23% overlap scenario in Figure 14. The overall performance without profile reduction is worse, but the outside profile alone is still better than combined, suggesting that the lesser performance of the combined profile *vs.* outside alone is due to bias in training the PLS coefficients (see above), or the "curse of dimensionality", rather than the bias in variable selection eliminating more outside models during PLS training.

Table 7. Fraction (number) of RFR models surviving variable reduction from the inside and outside profiles for NVS and GCCS at thresholds of r2>0.2, 0.05, and 0.

Co. / Cutoff	NVS		GCCS	
	inside	outside	inside	Outside
$r^2>0.20$ (2467)	0.80(345)	0.20(126)	0.45(191)	0.55(340)
$r^2>0.05$(3338)	0.69(1286)	0.31(623)	0.37(725)	0.63(1374)
$r^2>0$(4333)	0.65 (6570)	0.35(3564)	0.35 (3564)	0.65 (6570)

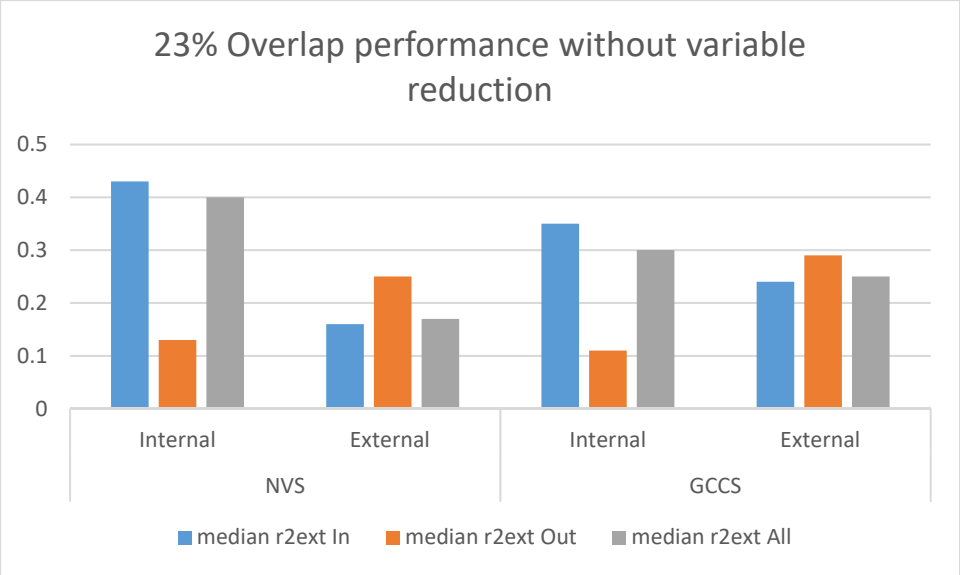


Figure 17. The performance of full-profile pQSAR models built using inside, outside, and combined (all) profiles on the internal and external test sets of compounds in NVS and GCCS under the 23% overlap collaboration scenario from Figure 14. Outside alone is still better than combined.

Consensus models

Outside profiles generally worked best for external compounds from competitor companies, and inside profiles worked best for the companies' own compounds. Thus, it appears that compounds are best predicted using the profile trained on more similar compounds. In a real collaboration, where only black-box models are shared, the user does not know whether the external compounds they want to predict are more similar to their own or their partners' compound collections. One option is to hedge bets with a consensus model. Pilot studies showed that consensus by taking the most active prediction from the 3 profiles worked poorly on both internal and external compounds. The "Consensus" bars of Figure 14-16, and rank-order plots in **Error! Reference source not found.** show that consensus by averaging predictions from the 3 profiles works very well, giving the best predictions on external compounds, and very comparable to the already outstanding performance of the inside profiles on internal compounds. Averaging predictions from the 3 profiles is therefore a safe strategy for compounds external to one's own corporate archive.

Practical issues of model sharing

There are several practical considerations for any multitask collaborative approach that shares partial models. Foremost is reverse engineering. RFR or other similarity-based models typically predict the highest activity for compounds similar to the most active training compounds. Generative or evolutionary chemistry algorithms might therefore be able to generate structures similar to a competitor's most active compounds by maximizing the single-task scores. Some other single-task modeling methods, such as ridge regression, should not share this liability. In addition, the shared, black-box models might be trained on linear combinations of individual assays to further obfuscate the training data while providing the same information to the PLS models.

We saw that collaborators who contribute just a few models benefit much more, at less cost, than collaborators who contribute many models. This "problem of the commons" could be alleviated by appointing an impartial broker, conceivably a robot, to distribute the single-task models. Each partner would receive only as many models from larger contributors as they contributed themselves. Even with such a broker, an unscrupulous partner might include bogus models or models from public data sets to inflate their contribution. Perhaps models could be trained to distinguish genuine from bogus models, or perhaps the partners could be audited to show they have data backing up their contribution. Of course, all multitask collaborative approaches that share models must address these issues.

An important caveat of this study is that meaningful overlap was only engineered between two large pseudo-companies. It is difficult to extrapolate how results will behave as more large companies are introduced. A large group of collaborators increases opportunities for overlap. However, the much larger profile might exacerbate the curse of dimensionality by increasing chance correlations and adding noise from irrelevant descriptors.

Conclusions

Multitask pQSAR is a natural fit for collaboration by sharing models, and therefore a practical method to explore the benefit of collaboration between companies. We attempted to use historical compound and assay data from the several companies that merged to form Novartis to evaluate pQSAR as a way for competitors to build collaborative multitask models by sharing black-box single-assay models without sharing compound structures, bioactivities or targets. Unfortunately, database annotations were not sufficient to assign reliably which compounds and

assays originated from which companies. As an alternative, “pseudo-companies” were constructed using compound identifiers, and assays were assigned based on which company’s compounds had been most extensively tested. This produced one very large NVS pseudo-company and 4 much smaller pseudo-companies. Unfortunately, this procedure left unrealistically little overlap between them.

For all 5 pseudo-companies with minimal overlap, pQSAR with just the inside profile dramatically improved internal predictions over the single-assay RFR models. In all cases, collaboration, *i.e.* addition of outside profiles, did not improve prediction on internal compounds even for small companies that received a huge number of NVS RFR models.

For the small companies, but not for much larger NVS, predictions on external compounds using outside or combined profiles was consistently better than inside profiles alone, and was always better than single-assay models (although not nearly as dramatically as for internal predictions). This supported the value of collaborative modeling for small companies, but not for a single, much larger one.

Subsampling NVS to create 5 similarly small companies greatly reduced the benefit of pQSAR for NVS although it still substantially outperformed single-assay RFR. Subsampling NVS also reduced collaborative pQSARs benefit for all companies’ external compound predictions, although collaboration was still beneficial, now for NVS as well.

As there was almost no overlap between the 5 pseudo-companies, and NVS was twice as large as the other 4 combined, the 4 smaller companies were merged into a single “GCCS” company, assays were moved from NVS to GCCS and moved compounds were assigned to GCCS to artificially introduce overlap. Predictions on external compounds improved as overlap increased. However, this artificial method of increasing overlap introduced complications: changing external compound sets, changing assays in the profiles and changing numbers of internal training compounds. Modifying the procedure to control for these artifacts did not change the qualitative improvement with overlap. Furthermore, the improvement was mainly in the moved, formerly NVS assays--either an increasing proportion of moved assays or increasing quality when the set of moved assays was fixed.

The inside profiles alone generally outperformed combined profiles for internal compounds, and outside alone outperformed combined for external compounds. The penalty for combining profiles might be due to PLS not finding important assay correlations due to the narrow applicability domains of the RFR models or to overtraining from an overabundance of descriptors.

Together these observations suggest that pQSAR predictions benefit external compounds that are similar to those used to train the outside models. In practice, one will not know if external compounds are more similar to one's own or one's collaborators' training sets. A consensus model using the average of predictions from all three profiles (in, out and all) was an effective solution.

These simulations show that sharing individual assay models among big pharma companies, without sharing compounds, targets or activity data, has the potential to broaden the applicability domain beyond that of each individual company's archives. Profile-QSAR is an ideal collaboration platform for its simplicity, efficiency, black-box safety, easy implementation with standard hardware and software, and relatively low cost. The bottom line is that *in all cases studied, multitask pQSAR was on average better than single-assay models--greatly benefiting predictions on internal compounds, and less dramatically benefitting external compound predictions. Collaboration never helped predictions on internal compounds, but did help predictions on external compounds, increasingly so with increasing compound collection overlap, but disappointingly, still far less than on internals.* Indeed, since collaborative pQSAR through model sharing is identical to pQSAR using actual shared data, and we know of no substantially better method for multitask modeling, we believe our conclusions should apply to collaborative modeling by any current method including even direct sharing of all the structures and assay data. We hope therefore that these simulations will serve as a preview of the benefits, requirements, and limitations of efforts among pharmaceutical companies to improve machine-learning predictions by sharing models or other methods of sharing data.

References

1. Verras, A.; Waller, C. L.; Gedeck, P.; Green, D. V.; Kogej, T.; Raichurkar, A.; Panda, M.; Shelat, A. A.; Clark, J.; Guy, R. K.; Papadatos, G.; Burrows, J., Shared Consensus Machine Learning Models for Predicting Blood Stage Malaria Inhibition. *J Chem Inf Model* **2017**, 57, 445-453.

2. Xu, Y. T.; Ma, J. S.; Liaw, A.; Sheridan, R. P.; Svetnik, V., Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *Journal of Chemical Information and Modeling* **2017**, 57, 2490-2504.
3. Bosc, N.; Felix, E.; Arcila, R.; Mendez, D.; Saunders, M.; Green, D. V.; Ochoada, J.; Shelat, A. A.; Martin, E.; Iyer, P.; Engkvist, O.; Verras, A.; Duffy, J.; Burrows, J.; Gardner, M.; Leach, A., MAIP: A prediction platform for predicting blood-stage malaria inhibitors. *J. Chem. Inf. Model.* **2020**, in reivew.
4. Bosc, N.; Felix, E.; Arcila, R.; Mendez, D.; Saunders, M.; Green, D. V.; Ochoada, J.; Shelat, A. A.; Martin, E.; Iyer, P.; Engkvist, O.; Verras, A.; Duffy, J.; Burrows, J.; Gardner, M.; Leach, A. https://www.ebi.ac.uk/chembl/maip/. https://www.ebi.ac.uk/chembl/maip/
5. Gedeck, P.; Skolnik, S.; Rodde, S., Developing Collaborative QSAR Models Without Sharing Structures. *Journal of Chemical Information and Modeling* **2017**, 57, 1847-1858.
6. Sheller, M. J.; Reina, G. A.; Edwards, B.; Martin, J.; Bakas, S., *Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation*. 2019.
7. IMI https://www.imi.europa.eu/projects-results/project-factsheets/melloddy. https://www.imi.europa.eu/projects-results/project-factsheets/melloddy (08/24),
8. MELLODDY MELLODDY Consortium Employs Federated Learning and Blockchain to Enhance AI Drug Discovery. https://astrixinc.com/melloddy-consortium-employs-federated-learning-and-blockchain-to-enhance-ai-drug-discovery/ (20oct2020),
9. Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J., Profile-QSAR: A Novel meta-QSAR Method that Combines Activities across the Kinase Family To Accurately Predict Affinity, Selectivity, and Cellular Activity. *Journal of Chemical Information and Modeling* **2011**, 51, 1942-1956.
10. Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C., Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, 57, 2077-2088.
11. Martin, E. J.; Polyakov, V. R.; Zhu, X. W.; Tian, L.; Mukherjee, P.; Liu, X., All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. *J Chem Inf Model* **2019**, 59, 4450-4459.