QUANTITATIVE MEDICINE

Transforming Drug Discovery™

Computational Research Engine™ Scientific Overview

Current Industry Practice and Limitations

Drug discovery and development is a long and costly process which begins with the identification of protein targets and ends after clinical trials [1]. These targets are often identified through basic science efforts. Many types of experimental assays have been developed to measure the effects of compounds on target proteins. In the early phases of drug discovery, compounds are tested using these assays to detect *desired* effects on the protein targets.

During the later phases of drug development, *undesired* effects from these compounds are often discovered but only after significant investment. Ideally, compounds would be chosen during the earliest phases of drug discovery using knowledge of the effects of screened compounds on many different targets while simultaneously selecting *for* desired effects and *against* deleterious effects.

To do this effectively, one would need to know the effects of every compound on every protein in the human body. Even with substantial automation, testing every compound against every available assay would be prohibitively expensive and time consuming. Even for a single assay, exhaustive testing of compounds is often very expensive.

Machine learning techniques have been used to reduce costs by predicting the effects of compounds *in silico*. When predicting the effects of compounds in a single assay, Quantitative Structure Activity Relationship (QSAR) analysis is frequently utilized [2, 3]. During a QSAR analysis, compounds are checked for the presence or absence of certain structural elements. The binary vector formed by checking for many different structures is called a "fingerprint."

By associating fingerprints with measured experimental outputs for only the assay of interest, a model can be built to predict the effects of other compounds in that assay. Using these models, decisions are made based on both the experimental observations and predicted effects of compounds. Typically, researchers used these results to identify compounds predicted to have desirable characteristics – e.g., low toxicity and high efficacy. A subset of these compounds is subsequently tested in the laboratory.

Computational Research Engine Innovations

Quantitative Medicine's Computational Research Engine™ (*CoRE™*) addresses three major shortcomings in the standard approach currently used by industry.

1. *Building predictive models from limited data.* When building predictive models using standard QSAR methods, typically only data from a single assay are used to learn the predictive models for each assay. This approach often limits the scope of accurate predictions to just the area of compound space where experimental data are available. With appropriate machine learning methods, data from many different assays can be used to improve the accuracy and scope of predictions for a single assay. Quantitative Medicine has curated a KnowledgeBase™ containing over 200 million experimental results from thousands of assays. Additionally, these historical experimental results are used in every

predictive model generated by *CoRE,* yielding much more accurate predictive models than standard approaches.

2. *Ineffective compound selection methods.* During the R&D cycle (Figure 1) and after generating predictions, those compounds that are predicted to have desirable characteristics are confirmed experimentally. The resulting data can be added to prior results and used to learn a new predictive model. By focusing only on those compounds predicted to be desirable, the value of the new data is limited. These models generally do not improve significantly with the iterative cycles of the R&D process. *CoRE* uses *active machine learning* to iteratively select compounds for testing which are expected to be the most informative. With this innovative approach, the accuracy of predictive models improves rapidly between cycles. As a result, highly accurate predictive models are generated and significantly fewer experiments are needed to accurately identify compounds with optimal characteristics.
3. *Ineffective stopping criteria.* It is important to determine when an identified compound is "good enough" to advance in the pipeline and whether better candidates are likely to exist. Experimental campaigns often continue long after the identification of optimal compounds. By using active machine learning methods, one can estimate the progress of a campaign and cease experimentation with confidence.

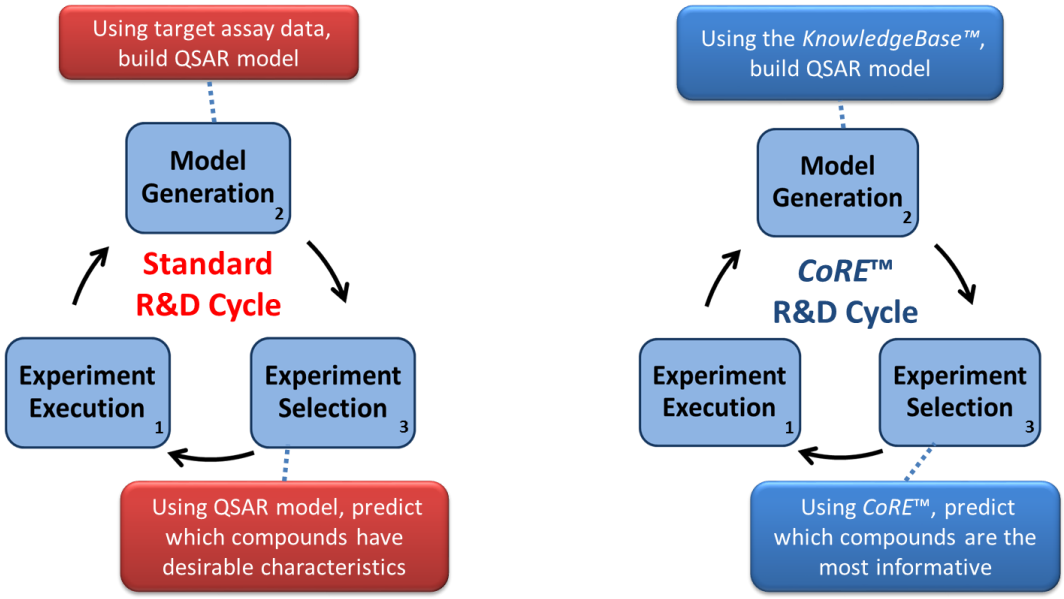

Figure 1: Primary differences between the standard R&D cycle and the *CoRE* R&D cycle

CoRE is utilized by constructing an experimental space composed of a set of assays and an *in silico* library of compounds that have been, or could be, synthesized and tested in those assays. Furthermore, adding data from previous campaigns will likely further improve predictive accuracy.

Clients populate the matrix with all available data and upload the data to *CoRE* by submitting an Excel spreadsheet. Any data can be utilized from all phases of drug discovery and development so long as the data were generated in experiments testing the effects of small molecules. Furthermore, *CoRE* does not require any descriptive information about the assays or their units of measurement.

CoRE processes the data, generates a set of predictions, and returns the predictions to the client along with the next batch of recommended experiments. Those experiments are then executed by the client and the resulting data are added to the Excel spreadsheet. This cycle is continued until the goals of the campaign are accomplished.

Quantitative Medicine has completed a number of proofs of value studies demonstrating the capabilities and benefits of *CoRE*. Results from a number of these studies are described below and are representative of expected results for future studies. Some proprietary details have been omitted.

Make Better Predictions Using *CoRE*'s KnowledgeBase™

The KnowledgeBase is curated from publicly available data. Included are over 200 million experimental results from more than 4,000 unique assays. Many methods of experimentation are represented such as high-throughput screening, high-content screening, *in vivo* screens and clinical trials. Data included from PubChem ranges from compound solubility to *in vivo* toxicity. Additionally, we have curated data from publications to improve the predictive scope and accuracy for specific toxicities such as mitochondrial toxicity.

CoRE considers all experimental data to be part of a very large matrix describing the universe of potential experiments. This matrix is illustrated in Figure 2; each row represents an assay and each column a compound. Thus, each point represents an assay-compound experiment. The colored points represent those experiments for which data are available from experimentation.

For each project run using *CoRE*, the client constructs an experimental space based on scientific and logistical constraints. This space is defined by the assays of interest and the compounds that could be synthesized and tested in those assays. An example client experimental space is shown by the green box. The meaning of the colors in Figure 2:

1) Green points represent data gathered by the client during the process of iterating through the *CoRE* R&D cycle.
2) Red points indicate data available in the KnowledgeBase.
3) Blue points represent data supplied by a client to increase the size and predictivity of the KnowledgeBase used by *CoRE* for only their project.

CoRE uses proprietary matrix completion algorithms developed specifically to function efficiently in matrices of very large scale. *CoRE* makes predictions for all unobserved experiments in the area of interest. Based on those predictions and active learning methods, *CoRE* selects a set of experiments to be executed for each iteration of the cycle.

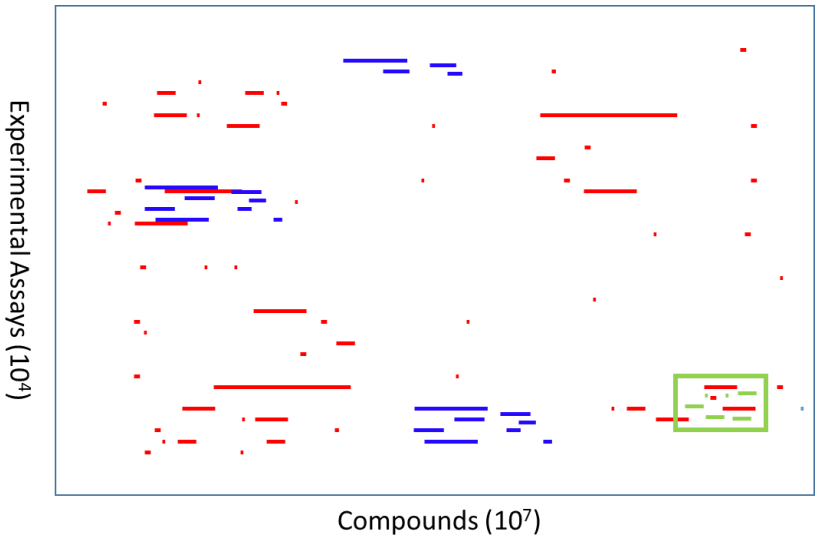

Figure 2: The experimental space is represented as a very large matrix. Each point in this matrix represents an experiment. If the point is colored, this indicates experimental data are available. To make predictions for the area of interest (the green box), data from the

KnowledgeBase (red) are combined with client data from prior campaigns (blue) and data from the active campaign (green).

Collaboration with Sanofi. Collaborators at Sanofi (Richard Brennan, Friedemann Schmidt, Richard Khan-Malek and Douglas Keller) identified various *in vivo* toxicity measurements from the Environmental Protection Agency's ToxCast Phase II project [4]. One of the major goals of the ToxCast project was to identify *in vitro* screening assays that are predictive of *in vivo* outcomes. To that end, more than 700 *in vitro* assays were run for more than 2,000 compounds. The toxicities of those compounds had been previously tested *in vivo*. The primary objective of the collaboration with Sanofi was to compare the predictive power of using different sources of information to predict the toxicities of interest:

1. compound fingerprints (standard QSAR) (FP) calculated with OpenBabel [5]
2. *in Vitro* screening data from ToxCast (TC)
3. *CoRE* KnowledgeBase (KB)
4. *CoRE* KnowledgeBase *and in vitro* screening data from ToxCast (KB+TC)

Richard Brennan identified 648 compounds that could be confidently labeled into the selected toxicity classes. He formed a training set by randomly selecting 75% of the compounds. The remaining compounds were withheld from the testing set. Only the training set was provided to Quantitative Medicine.

Using that training set to build predictive models with each source of information described above, we then predicted the results for the held out testing set. To measure model performance, the area under the ROC curve was measured for each toxicity using each information source. A summary of the predictive model results are shown in Table 1.

In vivo Toxicity Dataset Summary				Predictive Model Performance Using Different Sources of Information (Area Under ROC curve)			
***In Vivo* Toxicity**	**Pos**	**Neg**	**% Pos**	**FP**	**TC**	**KB**	**KB+TC**
Kidney Tubule Toxicant (necrosis, degeneration, regeneration, vacuolization, dilation)	52	310	14%	0.50	0.49	**0.60**	0.59
Hepatocellular Necrosis	115	230	33%	0.69	0.60	0.75	**0.76**
Hepatocellular Necrosis + ALT Increase	61	207	23%	0.70	0.74	0.76	**0.79**
Hepatocellular Necrosis + ALT + Bilirubin Increase	17	146	10%	0.58	0.50	**0.78**	0.66
Liver Toxicant (necrosis, apoptosis, degeneration, vacuolization, regeneration)	224	67	77%	0.67	0.59	**0.71**	0.58
Bile Duct Hyperplasia	56	225	22%	0.68	0.68	0.72	**0.82**
Systemic Necrosis (Internal organs excluding GI tract)	167	146	53%	**0.66**	0.51	0.61	**0.66**

Table 1: Predictive models were learned for seven toxicities described above using different sources of data: compound fingerprints (FP), i*n* vitro screening data from ToxCast (TC), *CoRE* KnowledgeBase (KB), *CoRE* KnowledgeBase and *in vitro* screening data from ToxCast (KB+TC). The quality of predictions for a test set was measured as the area under the ROC curve. For each *in vivo* toxicity, the performance of the best information sources has been bolded.

In this collaboration, it was concluded that using the KnowledgeBase improved prediction quality for all of the *in vivo* toxicities. For six of the seven toxicities (the seventh was a statistical tie), the best performing model used the KnowledgeBase. In four of the seven, the KnowledgeBase combined with ToxCast data provided the best model. This indicates *CoRE* can effectively use data from multiple sources to improve predictive accuracy. The primary reason for this improvement: by utilizing the KnowledgeBase, *CoRE* was able to find information in previously executed experiments which proved useful for making these predictions. Given the breadth of experimental assays included in the KnowledgeBase, these results can likely be extended to include many other types of experimental measurements.

After learning these models, we analyzed the models trained using KnowledgeBase data (denoted KB and KB+TC in Table 1) and estimated which assays from the KnowledgeBase contributed the most to the predictive power by assessing final model parameters. It was hypothesized that these assays might be predictive of the *in vivo* toxicities, but the predictivity of these assays were not assessed experimentally. The assays predicted to be most predictive are shown in Table 2.

These data suggest these relatively inexpensive *in vitro* assays could be used as an early inexpensive filter prior to expensive *in vivo* experimentation. *CoRE* can utilize a similar process for any expensive experimental method.

***In Vivo* Toxicity**	**Assay Source**	**Assay Description**
Bile Duct Hyperplasia	PubChem Assay 653515	Chemical Stability Assay
Systemic Necrosis	PubChem Assay 654360	Inhibition of T4 bacteriophage DNA polymerase
	PubChem Assay 654177	Screen for anti-amyloidogenic compounds
	PubChem Assay 2830	Screen for ABC transporter inhibitors
	PubChem Assay 1022	Screen for Mcl-1/Noxa interaction inhibitors
	PubChem Assay 1046	Screen for compounds affecting thrombin function
Hepatocellular Necrosis + ALT + Bilirubin Increase	PubChem Assay 1233	Screen for Inhibitors of Class II HMG-CoA Reductases
	PubChem Assay 485374	inhibitors of SUMOylation
	PubChem Assay 654167	Inhibition of rat brain monoacylglycerol lipase
Hepatocellular Necrosis + ALT Increase	PubChem Assay 654177	Screen for anti-amyloidogenic compounds
	PubChem Assay 654360	Inhibition of T4 bacteriophage DNA polymerase
Hepatocellular Necrosis	PubChem Assay 654177	Screen for anti-amyloidogenic compounds
	PubChem Assay 654360	Inhibition of T4 bacteriophage DNA polymerase

Table 2: Predictive models for each of these toxicities were analyzed to determine which, if any, PubChem assays were predictive of the *in vivo* toxicity.

The top predictive assays for these in vivo toxicities are shown as well as a description of the assay. The assays found to be relevant in the predictive models can also be used to suggest new avenues for research into the mechanism for a specific experimental measurement.

These results indicate using large amounts of data from both the KnowledgeBase, and as provided by a client, can improve the accuracy of predictive models accuracy for various toxicities. Currently, much of the research and development in drug discovery occurs in organizational silos, often with very little communication of knowledge with others who could benefit. These silos exist between organizations or even laboratories within those organizations.

Quantitative Medicine believes better transfer of scientific knowledge between and amongst these silos can enable more efficient discovery. It is challenging to make good decisions about experimental efforts when large amounts of data are available from diverse sources and experimental modalities. *CoRE* provides a very practical and efficient way of using data from multiple sources effectively. Furthermore, these results demonstrate the value of a client supplying data from prior campaigns (and even across various disease states) to further improve predictive accuracy - even if the relationship between the data and the current campaign is not readily apparent.

Efficiently Direct Experimentation Using the Computational Research Engine™

The solution to overcoming the current limitations of drug discovery and development techniques involves two major components: (1) building accurate and robust predictive models for the effects of many compounds on numerous targets; and, (2) using *active* machine learning to direct iterative experimentation and efficiently improve and effectively use these predictive models.

Active learning is a machine learning technique which is used to iteratively select experiments for execution by estimating which experiments will yield the most informative results. For a review of active learning, please refer to [6]. The computational methods used in the active learning process of *CoRE* are based around active learning research from Carnegie Mellon University. This research demonstrated active learning methods can efficiently direct experimentation to learn accurate predictive models for large systems of experiments [9, 10]. Since these publications, the methods have been significantly improved to scale to current capabilities as well as use the *CoRE* KnowledgeBase effectively.

Demonstration Technique. To test the active learning process, simulations were executed using data from collaborators. During these simulations, all available experimental results were hidden from the active learning process as if the experiments had never actually been executed. Once the active learning process requested specific experimental results, those results were revealed as if the experiments were executed. These new experimental results could be used to relearn a predictive model and continue the cycle until the end of the simulation. Each simulation was then assessed based on the specific criteria we were interested in exploring for that project.

ToxCast Phase 1 Active Learning Simulation. The objective was to determine how rapidly *CoRE* could learn an accurate predictive model for a relatively large experimental space. The experimental space was comprised of 1197 experimental assays tested across 309 compounds in the Phase 1 of the EPA ToxCast project [4]. In every round of each simulation, the results of all unobserved (unrevealed) experiments were predicted. The accuracy of these predictions was measured as the area under the ROC curve. For each round, 1% of all experimental results were selected by *CoRE,* and subsequently revealed to *CoRE,* for use in the model building process. Twenty simulations were run using different initial starting sets of experiments. The mean and standard deviation of the area under the ROC curve are shown as a function of the percentage of experimental space explored by the cyan line in Figure 3.

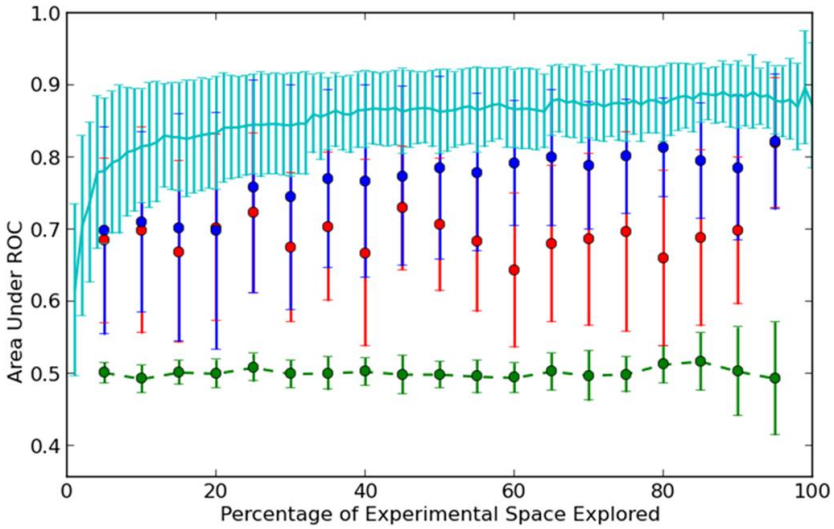

Figure 3: The mean and standard deviation of the accuracy of predictions (as measured by area under the ROC curve) as a function of the percentage of the experimental space explored is shown for four different methods of selecting and predicting missing data from the ToxCast Phase 1 dataset.

One standard approach to building a predictive model for an experimental space composed of small molecules is to select and test a single set of structurally diverse compounds. To simulate this process, compounds were clustered based on their binary fingerprint similarity. In increments of 5%, the most diverse X% of compounds were chosen and their results revealed.

The remaining unobserved (unrevealed) experimental results were predicted using RandomForest [11] (red) and Lasso regression [12] (blue) and the accuracy was measured. These processes are not typically iterative, so for each increment, selections were made independently of the other selections at different increments. A baseline random prediction with random selection was simulated and is depicted in green.

Using active learning to select experiments iteratively could yield models which were in expectation more accurate than the standard approaches using diversity-based selection. *CoRE* models generated using only 10% of the data were as accurate as the standard approach using 80% of the experimental space. This is because active learning processes (such as those used by *CoRE)* identify the informative experiments and ignore uninformative experiments. Simply finding the most diverse set of compounds, does not actually guarantee that you will find the most informative compounds. In this case, most of the information was gathered in the first 15% of the experiments. The remainder of the experimental results offered only small improvements in predictive accuracy. Significant savings of both time and cost could be realized by avoiding redundant and uninformative experimentation.

Optimization Campaign for Client 1

Quantitative Medicine was supplied with data containing the results of a completed optimization campaign. Researchers were attempting to find a compound with desirable characteristics across four assays - high stability, low *in vitro* toxicity, high binding affinity to confidential target, high *in vitro* efficacy. During this campaign, 80 compounds were synthesized and tested in all four assays. Prior to running any tests, the "best" compound was identified based on the desirable assay profile. The process of actually directing the campaign to identify the best performing compound was simulated using different selection strategies.

The following selection methods were used:

1. ***Random selection***: compounds were chosen randomly
2. ***Greedy selection***: For every batch, a model was built to predict the experimental results of unobserved compounds. Compounds were chosen which were predicted to *best match the desired profile*.
3. ***Active learning selection***: Use active learning to choose the most informative compounds to test.
4. ***CoRE* selection**: In the context of optimization, *CoRE* uses a hybrid method of greedy selection and active learning selection to select compounds such that model accuracy improves in each round; and, the increasingly accurate predictive models are used to select the best compounds.

In each simulation, three compounds were randomly selected for the initial round. In each of the following rounds, the designated selection method was used to select a batch of three compounds. A total of 250 simulations were performed for each selection method. After completing each simulation, we determined the batch number in which the "best" compound was identified. The mean and standard deviation for the number of batches required to find the best compound for each selection method are shown below in Figure 4.

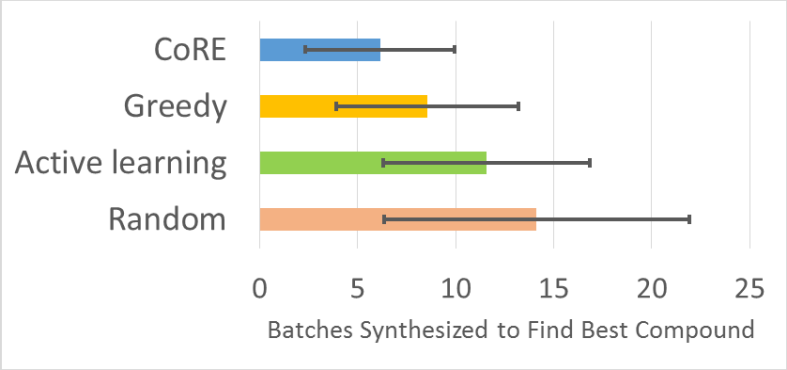

Figure 4: An optimization campaign was simulated 250 times with four different compound selection strategies. For each simulation, the number of batches to find the best compound was calculated. The mean and standard deviation for each selection strategy are shown.

CoRE selection was found to be the most effective of all selection methods for identifying the best compound in the fewest rounds. Random selection performed as would be expected and with very wide variation. Active learning selection performed slightly better than random.

With active learning selection, compounds were selected because they were expected to be informative and *not because they were expected to match the desired profile*. This means the best-matching compound in the clients' data was predicted to be somewhat informative - causing it to be selected more quickly than random selection.

With greedy selection, compounds were selected because they were predicted to match the desired profile. As the greedy selection simulations progressed, the models were not improving substantially from round to round. However, even with relatively poor models, the greedy approach was able to find the optimal compound in fewer batches than random selection or active learning selection.

CoRE selection was most successful because its efforts were focused on selecting promising and informative compounds. The effect of *CoRE*'s methodology is that the accuracy of the predictive model was improving in each round from the active learning selections. The more accurate model was effectively used in selecting compounds matching the desired profile.

Optimization Campaign for Client 2

A second collaborator constructed a dataset containing the results of an optimization campaign. Researchers were attempting to find a compound with desirable characteristics across three assays - low *in vitro* toxicity in two assays and high *in vitro* efficacy. During this campaign, they had synthesized and tested 185 compounds in all three assays.

Unlike the prior simulations for client 1, the best compound was not identified for Quantitative Medicine prior to running the simulation. The collaborator maintained control over the data - meaning experimental data was made available to *CoRE* only after the results were specifically requested by *CoRE*. This reduced the risk of potential training set contamination, but the resulting additional time to manage the data meant there was only time for two simulations to be completed.

In both of these simulations, *CoRE* used the KnowledgeBase to improve the accuracy of predictions. As the value of the KnowledgeBase had been previously demonstrated, the objective of this project was to further demonstrate the value of the active learning capabilities of *CoRE*.

In the first simulation, *CoRE* selection was used to select compounds expected to be informative (active learning) and predicted to match the desirable profile (greedy selection). In the second simulation, *CoRE* was only allowed to make greedy selections based on the KnowledgeBase predictions. Both simulations were initialized with the same set of 40 compounds. In each following round, 10 compounds were selected. After both simulations were complete, the batch number containing the "best" compound was identified.

The results obtained, as well as those achieved in the actual campaign, are found in Table 3. *CoRE* was able to identify the best compound in three rounds fewer than both the original campaign and the greedy approach. Again, this was primarily the result of *CoRE* improving its predictive model in each round and then leveraging the improved model to make better predictions.

Selection Method	Batch Best Compound Selected	Stopping Rule Invoked
Greedy selection	10	13
Actual Campaign	10	16
CoRE	**7**	**8**

Table 3: For each method of compound selection, we have shown the round in which the best compound was identified. Actual campaign performance is included as well for comparison to the researchers' original efforts. We also show the round the stopping rule was invoked.

In order to make these simulations even more insightful, our collaborators asked us to determine whether to continue each of the campaigns after each iteration, just as was done during the actual execution of the original campaign. We had to use the available data to determine the confidence with which the best compound was found in the dataset.

In the context of active learning, experimentation should generally be stopped when the accuracy of the predictive model no longer improves significantly from round to round. Unfortunately, it is impossible to effectively measure the accuracy of a model in the entire compound space when most experimental results are unavailable. As a replacement for model accuracy, model stability can be measured from round to round.

Stability was measured by assessing the extent to which model predictions changed from round to round. Once the models stabilized and no new "best" compounds were discovered the process was stopped.

Using this approach, the *CoRE* selection campaign was stopped after round 8. The greedy selection campaign was stopped after 13 rounds. The actual campaign performed by the client was stopped after 16 rounds. Because *CoRE*'s selection methods are focused not only on assessing and improving the model, but also on making accurate predictions, *CoRE* was able to use the information gained during this process to effectively stop the simulation prior to initiating a large number of extraneous experiments.

After completion of the simulations, we investigated how well model stability functioned as a suitable replacement for model accuracy. After both simulations were finished, our collaborators supplied the complete dataset for each optimization campaign. In each round of simulations, the predicted ranking of compounds was compared to the real ranking of compounds thus calculating a "ranking error". A higher ranking error indicated the predicted ranking of compounds was very different from the actual ranking.

Figure 5 below depicts the ranking error plotted as a function of the round of experimentation - for both simulations. Also depicted are markers indicating at which round of selection the best compound was discovered and when the campaign was terminated.

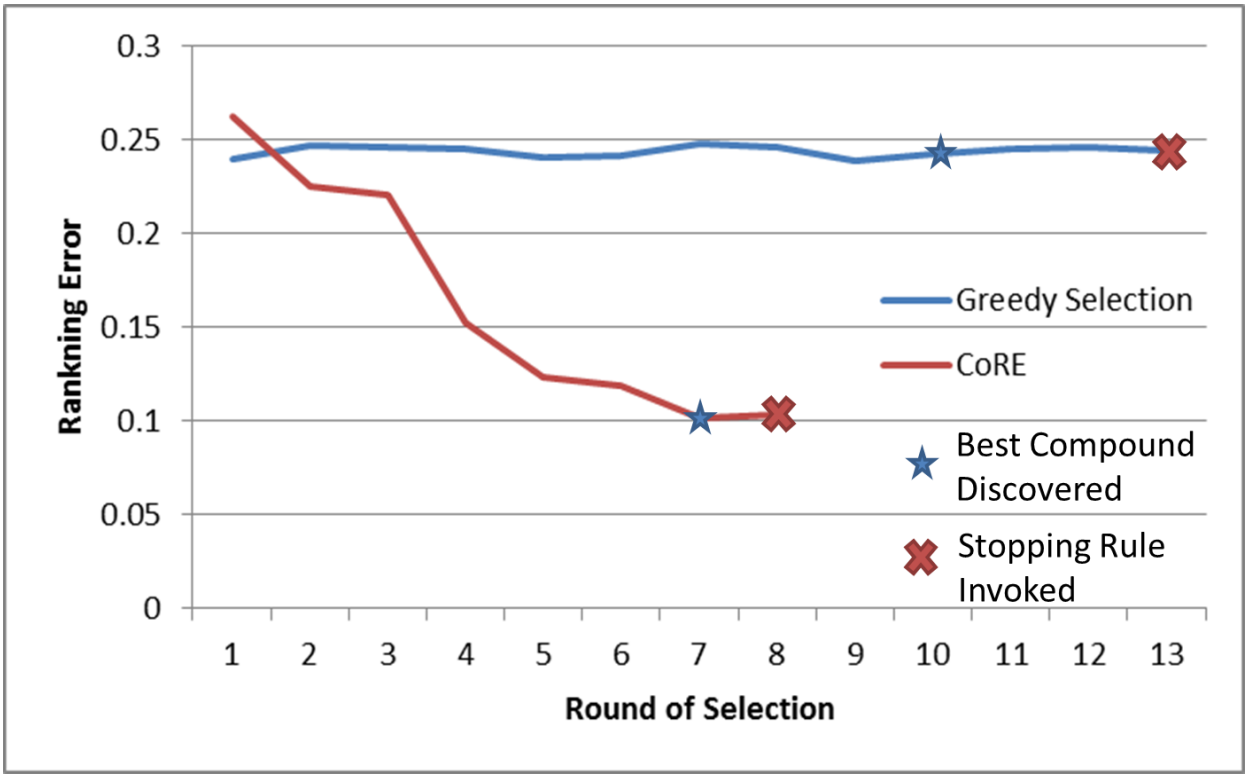

Figure 5: After the simulations were completed, the compounds' predicted ranks at each round were compared to the actual ranks. The difference between predicted and actual rank is shown as a function of the round of selection. Additional milestones for each campaign are marked as well.

The stopping rule worked very well for the *CoRE* selection simulation. Ranking error steadily decreased until round 7, the change was minimal from round 7 to round 8. This small change in error also corresponded to an observed small change in model predictions -- indicating the models had stabilized. Additionally, no new "best" compounds were discovered in round 8 so the simulation was terminated.

In the simulation of greedy selection, the ranking error never went down significantly and the predictions were not stable enough to end the simulation until round 13. Comparison of the greedy selection simulation and the *CoRE* simulation is striking. In the greedy selection simulation, compounds were never selected because they were expected to be informative and consequently the model did not improve substantially. In the *CoRE* selection simulation, some experiments were run because they were expected to be informative and as such the accuracy of the model improved significantly as the simulation progressed.

In practice, *CoRE* provides a wealth of information during each round permitting determination of when to stop a campaign with confidence. Quantitative Medicine expects the specific rules used by clients will change significantly depending on campaign- specific requirements and tolerance for risk. For example, more conservative clients may prefer the models maintain a stable state for three consecutive rounds of experimentation before stopping a campaign.

Conclusion

These results demonstrate that Quantitative Medicine's Computational Research Engine can significantly improve the productivity of drug discovery and development.

The *CoRE* KnowledgeBase provides an efficient and effective way of using large sources of information to generate useful predictions for drug discovery tasks. This unique capability enables *CoRE* to function as a central hub of information. This hub incorporates data from every campaign within an organization. The organization and its partners including CROs, academia and consortia can leverage the hub for great benefit. Although it was not demonstrated directly in these studies, one can also infer that the accurate models developed by *CoRE* can be used to predict common sources of late stage attrition very early in a campaign.

These methods are not limited to drug discovery and could be easily applied to other areas of biomedical research working in a system where multiple measurements could be made in the presence of multiple perturbations.

References

[1] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nature Reviews Drug Discovery,* vol. 9, pp. 203-214, 2010.

[2] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley and R. P. Sheridan, "Chemical Similarity Using Physiochemical Property Descriptors," *J. Chem. Inf. Comput. Sci.,* vol. 36, no. 1, pp. 118-127, 1996.

[3] R. P. Sheridan, M. D. Miller and D. J. K. S. K. Underwood, "Chemical Similarity Using Geometric Atom Pair Descriptors," *J. Chem. Inf. Comput. Sci.,* vol. 36, pp. 128-136, 1996.

[4] D. J. Dix, K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer and R. J. Kavlock, "The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals," *Toxicological Sciences,* vol. 95, no. 1, pp. 5-12, 2007.

[5] [Online]. Available: http://openbabel.sourceforge.net/. [Accessed 10 November 2010].

[6] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[7] Y. Fujiwara, Y. Yamashita, T. Osoda, M. Asogawa, C. Fukushima, M. Asao, H. Shimadzu, K. Nakao and

R. Shimizu, "Virtual Screening System for Finding Structurally Diverse Hits by Active Learning," *Journal of Chemical Information and Modeling,* vol. 48, pp. 930-940, 2008.

[8] M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta and C. Lemmen, "Active Learning with Support Vector Machines in the Drug Discovery Process," *Journal of Chemical and Informatio and Computer Sciences,* vol. 43, pp. 667-673, 2003.

[9] J. D. Kangas, A. W. Naik and R. F. Murphy, "Prediction of Biological Responses Using Protein and Compound Features and their Discovery using Active Learning," *BMC Bioinformatics,* vol. 15, p. 143, 2014.

[10] A. W. Naik, J. D. Kangas, C. J. Langmead and R. F. Murphy, "Efficient Modeling and Active Learning Discovery of Biological Responses," *PLOS One,* vol. 8, p. e83996, 2013.

[11] L. Breiman, "Random Forests," *Machine Learning,* 2001.

[12] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 58, pp. 267-288, 1996.

Contact

Joshua D. Kangas PhD
Founder / Chief Scientific Officer
JOSH.KANGAS@QTMED.COM
412.616.0641

QUANTITATIVE MEDICINE
Transforming Drug Discovery™