Realizing private and practical pharmacological collaboration

HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2019 May 15.

Published in final edited form as:

Science. 2018 October 19; 362(6412): 347–350. doi:10.1126/science.aat4807.

Realizing private and practical pharmacological collaboration

Brian Hie[*,1], **Hyunghoon Cho**[*,1], and **Bonnie Berger**[**,1,2]

[1]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA

[2]Department of Mathematics, MIT, Cambridge, MA 02139, USA

Abstract

While combining data from multiple entities could power life-saving breakthroughs, open sharing of pharmacological data is generally not viable due to data privacy and intellectual property concerns. To this end, we leverage modern cryptographic tools to introduce a computational protocol for securely training a predictive model of drug-target interactions (DTI) on a pooled dataset that overcomes barriers to data sharing by provably ensuring the confidentiality of all underlying drugs, targets, and observed interactions. Our protocol runs within days on a real dataset of more than a million interactions, and is more accurate than state-of-the-art DTI prediction methods. Using our protocol, we discover novel DTI that we experimentally validated via targeted assays. Our work lays a foundation for more effective and cooperative biomedical research.

One-sentence summary:

A computational protocol enables collaborative drug discovery by securely combining private pharmacological data.

Collaborative efforts to develop new, life-saving drug therapies have recently begun to take shape among pharmaceutical companies and academic labs, despite the highly competitive nature of the industry (1–5). Driving this transformation is the stalled or declining productivity of existing drug development pipelines amidst growing financial and regulatory pressures. Many in industry and academia are realizing that the difficult task of identifying novel drug candidates would be more successful if pooled experimental datasets and knowledge that go beyond any single organization are leveraged (6–8).

Until now, however, such forms of collaboration (1–3), including open-access data sharing partnerships like the Structural Genomics Consortium (https://www.thesgc.org), have been of limited scope because pharmacological data sharing is fundamentally restricted by

**Corresponding author: bab@mit.edu.

*These authors contributed equally to this work.

Author contributions: B. Hie, H. Cho, and B. Berger developed the computational methods. B. Hie and H. Cho implemented the protocol and performed the analyses. B. Berger supervised the research. All authors wrote the manuscript.

Competing interests: The authors declare no competing interests.

Data and materials availability: Code and data are available at http://secure-dti.csail.mit.edu. Drug target interaction datasets were obtained from DrugBank version 3.0 (https://www.drugbank.ca/) and STITCH version 5.0 (http://stitch.embl.de/). Protein domain information is available from the Pfam database (https://pfam.xfam.org/).

concerns about intellectual property and other financial interests. Currently, entities have to moderate the amount of data they share in order to maintain the confidentiality of drugs under development or the set of potential targets being tested, both of which may contain sensitive information about underlying research or business strategies.

Modern cryptography offers techniques to broaden pharmacological collaboration by greatly mitigating privacy concerns associated with data sharing. Secure multiparty computation (MPC) protocols (9) allow multiple entities to compute over their private datasets without revealing any information about the underlying raw data, except for the final computational output. Unfortunately, the promise of privacy-preserving collaboration has been severely hindered by the inability of existing secure computation frameworks to scale to complex computations over large datasets. In particular, analyzing a large amount of experimental data to predict new therapeutic interactions typically involves sophisticated algorithms that present a major computational challenge for MPC.

Here, we introduce a proof-of-concept, end-to-end pipeline for collaborative drug-target interaction (DTI) prediction based on secure MPC. Conceptually, our protocol divides computation across collaborating entities while ensuring that none of the entities has any knowledge about the private data (Fig. 1). We achieve this using a cryptographic framework known as secret sharing (10) in which a private value (“secret”) is collectively represented by multiple entities. Each entity is given a random number (“share”) in a finite field (i.e., integers modulo some prime number p) such that the sum of all shares modulo p equals the secret. Any strict subset of entities cannot extract any information about the underlying secret using their shares. Various protocols have been developed for performing elementary operations (e.g., addition or multiplication) over secret-shared inputs (10, 11), which taken together form the building blocks for a general purpose MPC framework.

Although secret sharing-based MPC typically requires overwhelming amounts of data communication between entities for complex and large-scale computations, very recent optimizations have leveraged techniques such as generalized Beaver triples and shared pseudorandom number generators to significantly reduce communication cost, enabling practical secure computation for challenging problems such as genome-wide association studies for a million individuals (12). Even with these advances, however, secure MPC is still infeasible for existing DTI prediction methods (13–16) primarily because their computations scale quadratically with the number of drugs (n) and the number of targets (m) in the dataset (e.g., n^2 or nm), which is prohibitive for realistic datasets with millions of compounds.

To achieve scalable computation while maintaining high accuracy, we draw from recent advances in deep learning (17) to train a neural network model for DTI prediction. Our neural network takes feature representations of a compound and a target as input, and predicts the interactivity of the given pair. Although we used chemical structure fingerprints and protein domain annotations as input features in our computational experiments (Methods), our framework readily generalizes to alternative features. We circumvent the quadratic complexity of existing methods by training our neural network over a dataset consisting of only the observed DTIs and a comparable number of putatively non-interacting

drug-target pairs, which is typically vastly smaller than the full drug-by-target matrix. Furthermore, we greatly reduce the cryptographic overhead of secure neural network training by optimizing our architectural choices for efficient MPC, such as using the rectifier (18) as our activation function and hinge loss as our loss function, both of which require only a single data-oblivious comparison to evaluate (Methods). These operations can be more efficiently implemented in MPC than alternatives such as the sigmoid function, which requires many such comparisons to accurately approximate. Taken together, our efficient protocol allows our neural network to securely train over a wide area network (WAN) in under four days on a dataset with more than a million training instances (Table S1). In contrast, a recently proposed protocol for privacy-preserving neural network training (19) requires months of communication time over a WAN to train on an image dataset of smaller scale (60k examples, 784 features).

To demonstrate the accuracy of our securely trained neural network for DTI prediction (Secure DTI), we compared it to state-of-the-art DTI prediction techniques, including those based on matrix factorization with side information (CMF) (14), network diffusion (NetLapRLS; BLMNII) (16, 20), and heterogeneous data integration (DTINet; HNM) (15, 21) on a standard benchmark dataset (22) with 1,923 observed interactions (Methods). In addition to newly ensuring the confidentiality of the pooled input data during the computation, Secure DTI surpasses the performance of all baseline plaintext methods in cross validation accuracy (Fig. 2a, S1), a surprising result in light of the optimizations we made to achieve practical scalability. Our improvement over the best-performing baseline (DTINet) is statistically significant (one-sided Wilcoxon rank-sum p-value of 0.006) and further pronounced in a more challenging but realistic evaluation setting where the entire interaction space is considered as the test data rather than a balanced number of positive and negative examples (Fig. S1f). None of the baseline methods can be efficiently implemented in MPC for the purposes of secure collaboration, due to their quadratic complexity in the number of drugs and targets. In contrast, matrix factorization without side information (MF) lends itself to efficient MPC (23) but at the cost of greatly diminished model performance (Fig. 2a, 2b, S1).

We next set out to demonstrate the scalability and predictive performance of Secure DTI on a much larger dataset that more accurately represents the scale of cross-institutional collaboration. We obtained 969,817 interactions from the STITCH 5 human dataset (24), to our knowledge the largest publicly available DTI dataset, and evaluated the cross validation performance of Secure DTI. Even on the challenging task of predicting DTIs of previously unseen compounds, Secure DTI achieved high accuracy (AUPR of 0.95), which substantially outperforms matrix factorization methods (AUPRs of 0.50 and 0.43; Fig. 2b, S2). Other baseline methods could not be reasonably applied to a dataset of this size (even in plaintext) due to their quadratic scalability. In contrast, Secure DTI took less than four days to train on millions of interactions over a WAN (Methods) and efficiently scaled with a linear dependence on the number of interactions in the dataset (Fig. 2c, Table S1). Even training on two million interactions, we extrapolate the total runtime for one epoch (one linear pass over the full, shuffled training set) to be around 2.2 days. In practice, we expect our model to achieve high accuracy in only a few training epochs; we obtained all our reported results after 1.5 epochs. Additionally, given that our protocol admits flexibility in

the choice of predictive model, we also securely trained a support vector machine (SVM) instead of a neural network; the SVM reduced predictive performance (Fig. S2).

To go beyond cross validation and demonstrate the potential for novel discoveries that can result from our collaborative pipeline, we trained Secure DTI on all STITCH 5 interactions and scored the remaining possible drug-target pairs for interactivity, which is closer to how our pipeline would be used in a real-world setting. We controlled for bias toward highly represented drugs and targets in the dataset by either (i) filtering out any prediction involving both a drug and target highly represented in the original dataset (Secure DTI-A) or (ii) sampling negative examples (i.e., non-interactions) during model training such that each drug or target was seen at the same relative frequency in the negative examples as in the positive examples (Secure DTI-B) (Fig. 2b). In both cases, many of our top predictions (5/12 for Secure DTI-A and 9/12 for Secure DTI-B) were validated by our own targeted assay experiments or by published experimental studies that have not yet been deposited into the STITCH database (Table 1), including a novel interaction between imatinib and ErbB4, for which we could not find any existing experimental support. The top prediction from both methods was an interaction between the estrogen receptor (ER) and droloxifene, which had reached phase III clinical trials as an ER modulator for advanced breast cancer (25). Similarly, the predicted interaction between the vitamin D receptor (VDR) and seocalcitol has been clinically well-established (26). Furthermore, some predictions without direct activity have strong evidence for an indirect functional interaction; for example, nutlin-3 has been shown to inhibit PARP1 protein levels through p53-dependent proteasomal degradation in mouse fibroblasts (27). All of our findings were obtained without revealing any information about the underlying drugs, targets, and interactions during the computation.

To provide enhanced interpretability of our reported results, we incorporated an additional step into our pipeline for securely evaluating the impact of individual input features on the prediction outcome (Supplementary Text). When applied to our top predictions from the STITCH database, this capability linked drugs to specific ligand-binding or functional sites within the predicted target (Table S3).

We envision a real-world scenario in which multiple participating entities contribute secret-shared data to train a machine learning model on a joint pharmacological dataset (Fig. 1). After training, the model can be made available to all participants, or the model could remain private such that entities receive a number of predictions commensurate with the amount of data they contribute in order to incentivize participation. As more training data will most likely result in greatly improved performance (Fig. S3, Supplementary Text), entities will be incentivized to share information in a way that is mutually beneficial and has provable privacy guarantees.

Our pipeline is secure under the “honest-but-curious” security model in which the collaborating entities follow the correct protocol and do not collude to reconstruct the data. This is a substantial improvement over the current state of biomedical research where privacy concerns hinder any collaboration, but our framework can also be extended to achieve even stronger security guarantees. As our security guarantee holds as long as at least one entity is honest during the main computation (Methods), we can relax the no-collusion

requirement by introducing additional collaborating entities into our protocol, which does not substantially increase total computation time but does increase communication linearly in the number of entities. If we require security against malicious entities who deviate from the protocol during the online computation, we can include a message authentication code (MAC) with each message. At the end of the protocol the MAC is verified to ensure that each step was performed according to the protocol specification, a technique introduced in the SPDZ framework (28). This approach roughly doubles computation and communication, offering a tradeoff between security and performance that can be adjusted according to specific study requirements.

Although our pipeline does not consider adding noise to the final computation output to limit information leakage, a technique known as differential privacy (29, 30), methods being developed for differentially private neural networks can be used in conjunction with our protocol (31). An alternative strategy for collaborative neural network-based prediction is to train local models in plaintext and to use secure protocols only when periodically averaging over these models, thus minimizing the amount of cryptographic overhead (32, 33). However, this approach is vulnerable to reverse engineering-based attacks in which a malicious collaborator jointly trains a local model (e.g., a generative adversarial network) that uncovers information about private data owned by honest collaborators, even when differential privacy techniques are applied (34). In contrast, securely training a single model over a decentralized network of computing parties, as in our pipeline, is not vulnerable to such attacks.

Our privacy-preserving protocols generalize to other large-scale data sharing problems beyond drug discovery, with the highest potential for impact in areas that suffer from a lack of collaboration due to privacy concerns, such as predictive analyses of electronic health records. Our practical demonstration of secure, large-scale machine learning with neural networks may also provide a useful blueprint for enhancing privacy in many other domains where neural networks have shown to be successful.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank E. Irvine and P. Macaluso for assistance with validation experiments.

Funding: B. Hie and H. Cho are partially supported by NIH grant R01GM081871 (to B. Berger). H. Cho is also partially supported by the Kwanjeong Educational Foundation.

References

1. Reardon S, Pharma firms join NIH on drug development. Nature (2014).
2. Levy J, The age of collaboration: why pharma companies now have to work together. Pharmafile (2015).
3. Wilhelm M, Big Pharma Buys Into Crowdsourcing for Drug Discovery. Wired (2017).
4. Hunter J, Collaboration for innovation is the new mantra for the pharmaceutical industry. Drug Discov. World (2014).

5. Johnson & Johnson Innovation Announces New Collaborations Advancing Ground-Breaking Biomedical Innovation Around the Globe. *Press Release* (available at https://www.jnjinnovation.com/sites/default/files/jji_bio_2017_press_release_06-15-17.pdf).
6. Khanna I, Drug discovery in pharmaceutical industry: Productivity challenges and trends. Drug Discov. Today. 17 (2012), pp. 1088–1102. [PubMed: 22627006]
7. Cressey D, Traditional drug-discovery model ripe for reform. Nature. 471 (2011), pp. 17–18. [PubMed: 21368796]
8. Paul SM et al., How to improve RD productivity: The pharmaceutical industry's grand challenge. Nat. Rev. Drug Discov 9 (2010), pp. 203–214. [PubMed: 20168317]
9. Cramer R, Damgård IB, Nielsen JB, Secure Multiparty Computation and Secret Sharing (Cambridge University Press, Cambridge, UK, ed. 1, 2015).
10. Ben-Or M, Goldwasser S, Wigderson A, Completeness Theorems for Non-Cryptographic Fault Tolerant Distributed Computation. Proc. 20th Annu. ACM Symp. Theory Comput., 1–10 (1988).
11. Catrina O, Saxena A, Secure computation with fixed-point numbers. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 35–50 (2010).
12. Cho H, Wu D, Berger B, Secure genome crowdsourcing for million-individual association studies. Nat. Biotechnol 36, 547–551 (2018). [PubMed: 29734293]
13. Cobanoglu MC, Liu C, Hu F, Oltvai ZN, Bahar I, Predicting drug-target interactions using probabilistic matrix factorization. J. Chem. Inf. Model 53, 3399–3409 (2013). [PubMed: 24289468]
14. Zheng X, Ding H, Mamitsuka H, Zhu S, Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '13, 1025 (2013).
15. Luo Y et al., A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Commun 8, 573 (2017). [PubMed: 28924171]
16. Xia Z, Wu L-Y, Zhou X, Wong STC, Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst. Biol 4, 1–12 (2010). [PubMed: 20056001]
17. LeCun Y, Bengio Y, Hinton G, Deep learning. Nature. 521, 436–444 (2015). [PubMed: 26017442]
18. Glorot X, Bordes A, Bengio Y, Deep sparse rectifier neural networks. AISTATS '11 Proc. 14th Int. Conf. Artif. Intell. Stat. 15, 315–323 (2011).
19. Mohassel P, Zhang Y, SecureML: A System for Scalable Privacy-Preserving Machine Learning. Proc. - IEEE Symp. Secur. Priv., 19–38 (2017).
20. Mei JP, Kwoh CK, Yang P, Li XL, Zheng J, Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics. 29, 238–245 (2013). [PubMed: 23162055]
21. Wang W, Yang S, Zhang X, Li J, Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics. 30, 2923–2930 (2014). [PubMed: 24974205]
22. Knox C et al., DrugBank 3.0: A comprehensive resource for "Omics" research on drugs. Nucleic Acids Res. 39 (2011), doi:10.1093/nar/gkq1126.
23. Nikolaenko V. Privacy-preserving matrix factorization; Proc. 2013 ACM SIGSAC Conf. Comput. Commun. Secur. - CCS; 2013. 801–812.
24. Szklarczyk D et al., STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 44, D380–D384 (2016). [PubMed: 26590256]
25. Buzdar A et al., Phase III randomized trial of droloxifene and tamoxifen as first-line endocrine treatment of ER/PgR-positive advanced breast cancer. Breast Cancer Res. Treat 73, 161–175 (2002). [PubMed: 12088118]
26. Tocchini-Valentini G, Rochel N, Wurtz JM, Moras D, Crystal Structures of the Vitamin D Nuclear Receptor Liganded with the Vitamin D Side Chain Analogues Calcipotriol and Seocalcitol, Receptor Agonists of Clinical Importance. Insights into a Structural Basis for the Switching of Calcipotriol to a Receptor . J. Med. Chem 47, 1956–1961 (2004). [PubMed: 15055995]
27. Matsushima S et al., An Mdm2 antagonist, Nutlin-3a, induces p53-dependent and proteasome-mediated poly(ADP-ribose) polymerase1 degradation in mouse fibroblasts. Biochem. Biophys. Res. Commun 407, 557–561 (2011). [PubMed: 21419099]

28. Damgård I, Pastro V, Smart N, Zakarias S, Multiparty computation from somewhat homomorphic encryption. Crypto. 7417 LNCS, 643–662 (2012).

29. Dwork C, McSherry F, Nissim K, Smith A, Calibrating noise to sensitivity in private data analysis. Theory Cryptogr., 265–284 (2006).

30. Simmons S, Sahinalp C, Berger B, Enabling Privacy-Preserving GWAS in Heterogeneous Human Populations. Cell Syst. 3, 54–61 (2016). [PubMed: 27453444]

31. Abadi M. Deep Learning with Differential Privacy; ACM Conf. Comput. Commun. Secur.; 2016. 308–318.

32. Takabi H, Hesamifard E, Ghasemi M, Privacy Preserving Multi-party Machine Learning with Homomorphic Encryption. Proc. Work. Priv. Multi-Party Mach. Learn., 1–5 (2016).

33. Shokri R, Shmatikov V, Privacy-preserving deep learning. 2015 53rd Annu. Allert. Conf. Commun. Control. Comput. Allert. 2015, 909–910 (2016).

34. Hitaj B, Ateniese G, Perez-Cruz F, Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning. arXiv (2017).

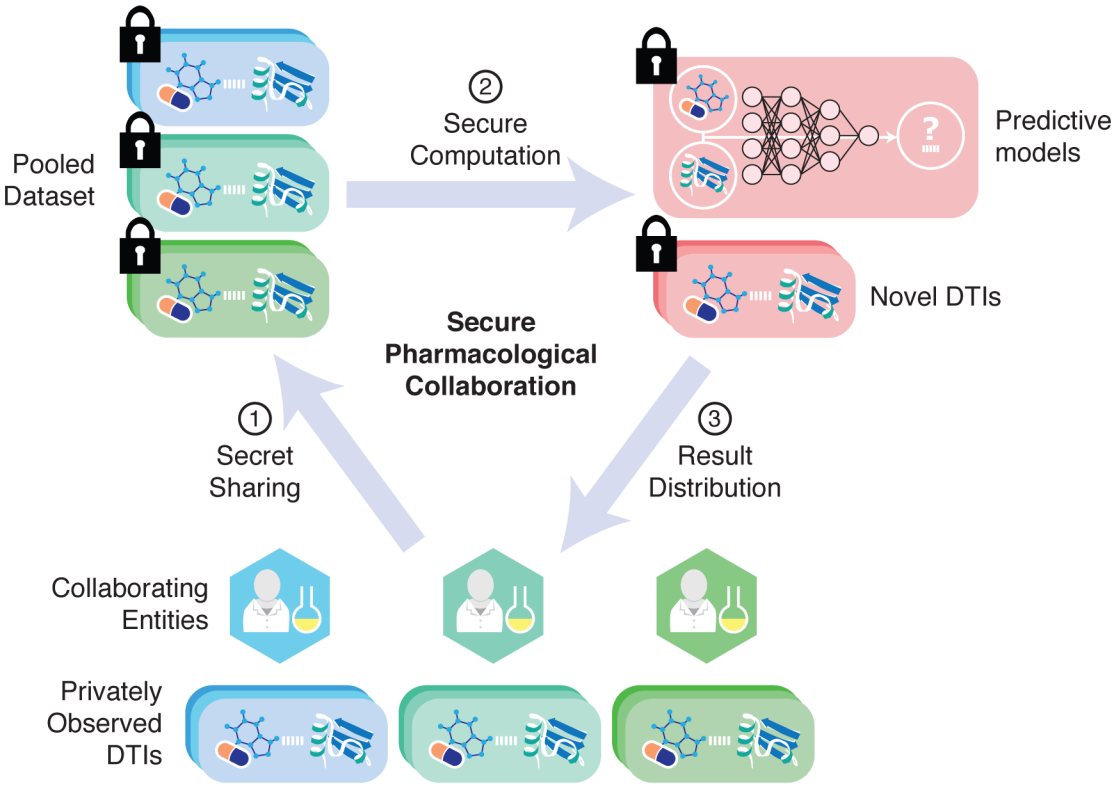

Fig. 1. Secure pipeline for pharmacological collaboration.

Collaborating entities (e.g., pharmaceutical companies or research laboratories) have large private datasets of drug-target interactions (DTIs), as well as corresponding chemical structures and protein sequences. In our protocol, the entities first use secret sharing to pool their data in a way that reveals no information about the underlying drugs, targets, or interactions (**Step 1**). The collaborating entities then jointly execute a cryptographic protocol that trains a predictive model (e.g., a neural network) on the pooled dataset (**Step 2**). The final model can be made available to participating entities, or be used to distribute DTI predictions to participants in a way that encourages greater data sharing (**Step 3**).

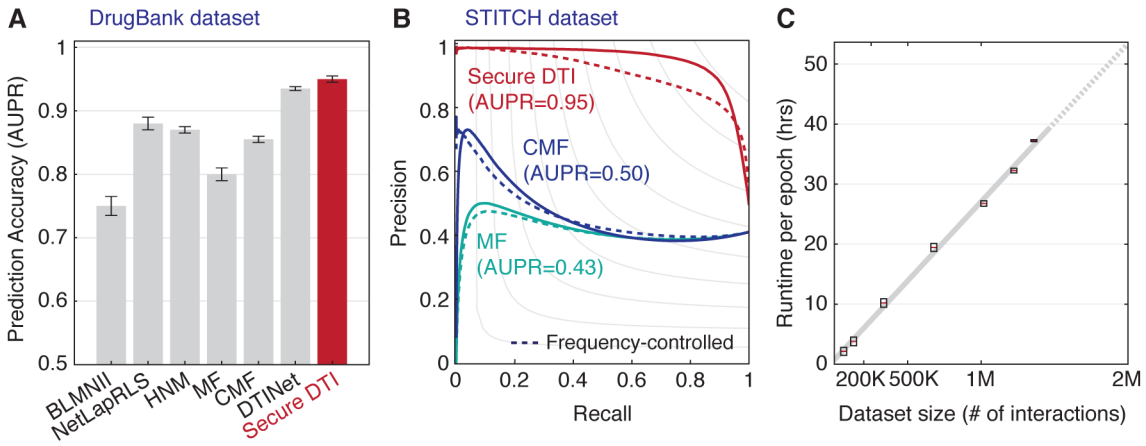

Fig. 2. Prediction of drug-target interactions.

(A) Predictions from the DrugBank 3.0 dataset. Bar height corresponds to mean AUPR (area under the precision-recall curve) and error bars correspond to the standard deviation. We compared Secure DTI to the plaintext methods BLMNII (20), NetLapRLS (16), HNM (21), MF (13), CMF (14), and DTINet (15) as reported in Luo *et al.* (15) by means of 10-fold cross validation (CV) on balanced training and test sets (Methods; see Fig. S1 for other evaluation settings). **(B)** Predictions form the STITCH 5 dataset with over a million drug-target pairs. Secure DTI is compared to matrix factorization with (CMF) and without (MF) side information (see Fig. S2 for other evaluation settings). Solid line, sampling negative examples randomly; and dashed line, sampling negative examples while matching the relative frequencies of drugs and targets to those in the positive examples, representing a more challenging test case. Reported AUPRs are for the solid curves. **(C)** Runtime of our training protocol, over a local area network (LAN), for different dataset sizes. Box height represents the standard deviation.

Table 1.

Predicted out-of-dataset drug-target interactions.

We trained Secure DTI on all human drug-target interactions from STITCH 5, which we used to score and rank all pairs of drugs and targets that are not in the STITCH database. We implemented two methods to control for model bias toward overrepresented drugs and targets, either **(a)** filtering out predictions involving a drug and target that are both highly represented in STITCH or **(b)** retraining Secure DTI such that the negative training examples had an equal representation of drugs and targets as the positive training examples (Methods). (*) indicates predicted interactions that were experimentally validated, including all testable interactions without existing literature support. Interactions labeled N/A involve commercially unavailable compounds and so could not be tested. We labeled an interaction as "active" if its IC50 was less than 100 μM, "weakly active" if there was close to 50% inhibition at 100 μM (our highest tested concentration), and "inconclusive" if activity was observed but only at few high concentration levels, a potential artifact of compound aggregation. References and additional information are given in Table S2.

	(a) Secure DTI-A			(b) Secure DTI-B		
Rank	**Drug**	**Target**	**Experimental Validation**	**Drug**	**Target**	**Experimental Validation**
1	Droloxifene	ERα	Active *	Droloxifene	ERβ	Active *
2	Droloxifene	ERβ	Active *	CHEMBL601690	p110α	Active
3	Imatinib	ErbB3	Inconclusive *	Droloxifene	ERα	Active *
4	Imatinib	ErbB4	Active *	Seocalcitol	VDR	Active
5	Nutlin-3	PARP1	Inactive *	AGN-PC-0A9TBG	PPARγ	Active
6	Droloxifene	PgR	Inactive *	CHEMBL589864	p110α	Active
7	Actinomycin D	PARP1	Weakly active *	T5958429	PARP1	Active
8	Hoechst 33258	PARP1	Inactive *	AGN-PC-0N7PYE	Factor Xa	Active
9	GW-501516	GR	Inactive *	AGN-PC-00DJ3O	PPARγ	Active
10	AGN-PC-0BFP0W	Lck	Active	AGN-PC-0NA8NJ	PTPRZ1	N/A
11	CHEMBL2332055	mGluR1	Inconclusive *	AGN-PC-0NA8NJ	PTPRG	N/A
12	CHEMBL2332055	mGluR5	Inconclusive *	AGN-PC-088DZ9	PROC	N/A

* Experimentally assayed in our study.