Title
Methods for Cheminformatic Prediction

Author
Caceres, Elena L

Publication Date
2021

Peer reviewed|Thesis/dissertation

Methods for Cheminformatic Prediction

by
Elena Caceres

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Michael Keiser — Michael Keiser
Chair

DocuSigned by:
Jason Gestwicki — Jason Gestwicki
DocuSigned by:
Brian Shoichet — Brian Shoichet
C8380A8942D641D...

Committee Members

Acknowledgements

I entered the Biological and Medical Informatics program at UCSF with a keen longing to learn more about how to structure scientific inquiry. From my start as a student with an incomplete and often second-hand knowledge of computer science, mathematics, and biology, I have tried to become what what some might consider to be an independent thinker with an incomplete, first-hand understanding of cheminformatics. Writing my thesis has cemented my journey on paper and I feel deeply indebted to all those along the way who made my path possible. I could not have done it alone. To you, reader, I am eternally grateful for the time you spend to read my work and for any kindness and support we have shared along the way; Thank you.

First, I would like to thank the family who made me, and whose innumerable members continued to love me as I proceeded to re-make myself over and over again. Mom, your interest in numbers, belief in mentorship, and commitment to leading by example has been a guiding light for my own professional development. Thank you for pushing me to ask for more and then challenging me to pursue it. Dad, gracias por su fe. Agradezco a ustéd que siempre me haya animado a trabajar más duro en las escaladas de montañas para sufrir menos. Además, gracias por apoyarme a pesar de mi español malo. Paul, thank you for your fierce support, for your help fixing literally anything, and for obligating me to always speak up on issues of right and wrong. Ultimately, science without a strong commitment to ethics decays toward public menace. Thank you for holding me accountable. Grammy and Grampy, Thank you for always encouraging me to ask "why" and for always bringing me

books at the holidays. I deeply appreciate your support for my academic interests from an early age. And to *all eight of my siblings*, thank you for encouraging me to pursue a sense of adventure. I would not have decided to "full send" a PhD without learning how to dive in head first in order to keep up with the herd. Each of you explore and improve the world around you and I am so proud to be your sister.

To my emotional support humans. This work was completed during a worldwide pandemic which continues to devastate millions. I am eternally grateful for all of your support over the last few years, but particularly for this past year. Thank you for all the phone calls. Janine Siegal, thank you for giving me trust in the beauty and struggle of re-invention. Thank you for rooting yourself firmly in my corner and for being one of my top three sisters. You are always there for me and I truly appreciate the kind and well-reasoned advice that I never take. Leanna Morinishi, you were the best recruitment roommate and hiking buddy that anyone could have asked for. You are just the warmest person and a fantastic scientist to boot. I am so thankful you were my friend from the start and I am honored that I could marry you (to your husband). Kaitlin Hulce, you are a steadfast friend, a shameless snack buddy, and my favorite running partner. Time spent with you is always a joy, and I am truly grateful for your gentle support, your earnestness, and your enthusiasm for bets. Lydia Ellul, thank you for making me feel welcome and accepted, for every single card, and for teaching me how to rely on knowing the data whenever I am unsure of myself.

To all my mentors , I would like to extend my gratitude for your time and effort to bring me into the scientific fold. Thank you to Ellen Potter and Dona Mapston for creating a high school summer program where I could learn to love science (even if there were negative results). Thank you to Ann Atkins, Han Cho, and Ron Evans for my first job and for teaching me how to pipette. Talitha van der Meulen, thank you for letting me follow you around the cell culture room. Your support for my transition from high school intern to college lab assistant was deeply formative in my decision to pursue a PhD. Thank you for teaching me so much about how to find joy in lab, how to talk to scientists, and how to

always be looking for new things to learn. Mark Huising, thank you for taking me on as your assistant. You taught me how to be independent in the lab and encouraged me to talk with you about experiments to try. Thank you for encouraging me to try new majors and for believing I could take my newly learned skills from my CS classes and apply them to our datasets. I doubt I would have discovered bioinformatics without your support.

I would also like to thank my friends who have stuck with me since childhood. Thank you Hailey Cunningham, you are the strongest person I know and I am grateful you could teach me how to be more assertive. Jennifer Oh, you have been a constant in my life. Thank you for being a great science lunch buddy and for your friendly ear. Miranda Stratton, as my Latina in science twin, you have been an inspiration and source of support over the years. We mirrored each other from grade school to graduate school and although we are finally taking different paths, I can't wait to see what you do next. I am so glad we reconnected.

To my community at UCSF, thank you for making our institution a fun place to come into work. Dear Julia Molla, Nicole Flowers, and Rebecca Dawson: making it through iPQB would not have been possible without all of your support and multiple follow-up e-mails. I am genuinely amazed at the amount of work it takes to herd professors and to get so many recalcitrant students through the finish line. Thank you. In addition to my thesis lab, I also had the pleasure to rotate through the labs of Joe DeRisi and Nadav Ahituv. You have my thanks for teaching me the knowledge I do have about genomics. Joe and Julia, thank you for finding me at SACNAS and telling me to apply. I would also like to thank all of my friends from iPQB and CCB for impromptu coffee and chats, but I would particularly like to draw attention to Alain Bonny, Allison Wong, Stephanie See, and Seth Axen. Graduate school would not have been the same without you. Thank you Meredith Kuo for your enthusiastic collaboration and for teaching me all about high throughput screening entirely over Zoom. Thank you to my committee members Jason Gestwicki and Brian Shoichet whose mentorship, scientific feedback, and professional guidance have been essential to my training. Thank you for your candor and for pushing me to learn more and letting me know

about my blind spots.

Toward the end of my PhD, I was incredibly lucky to intern with some fantastic scientists at Merck. Thank you to the entire Computational and Structural Chemistry team especially Song Yang, Essam Metwally, Chen Cheng, and Scott Hollingsworth for your input and help starting up a summer project. Alan Cheng, thank you for taking a chance on me and taking me on for an internship. You pushed me to try new approaches and I am continually amazed at how easy it is to learn from you. I am deeply appreciative of your structured guidance, enthusiasm, and large knowledge base.

Thank you to all of the members of the Keiser Lab past and present for making the lab a fun and enjoyable place to spend time. I really enjoyed all of the the scientific discussions, gym excursions, jumping contests, and practical jokes. Thank you to Garrett Gaskins for recruiting me, giving me so much advice, and eventually admitting that you like the term BFFL. Thank you Kangway Chuang for your scientific rigor, your deep knowledge of chemistry/really anything else in science, and for holding the lab accountable. Laura Gunsalus, thank you for being an agent of change and for renewing my interest in graphs. Jessica McKinley, I'm happy to have you as my science buddy and for your constant encouragement and support. Nicholas Mew, thank you for being a fantastic coworker and collaborator. Sahru Keiser, thank you for your help with scheduling. You are a gem.

And finally, to my PI and scientific advisor, Michael Keiser. My thesis would not have been possible without you. Thank you for admitting me into your fledgling lab, and for all of your advice, patience, and guidance over the years. You have taught me so much about how to speak to an audience and I am a better scientific communicator for having been your student. Adding to the scientific record can take time and feels like emptying an ocean with a thimble. Thank you for giving me the opportunity and room to grow as a scientist and a person. I am grateful you made a lab filled with encouraging and supportive people from whom I could learn how to fill my own thimble in uncharted waters.

Contributions

Chapter 1

Cáceres, E. L., Mew, N. C. & Keiser, M. J. Adding Stochastic Negative Examples into Machine Learning Improves Molecular Bioactivity Prediction. *Journal of Chemical Information and Modeling* **60.** PMID: 33245237, 5957–5970. doi:10.1021/acs.jcim.0c00565 (2020)

Chapter 2

Axen, S. D., Huang, X.-P., Cáceres, E. L., Gendelev, L., Roth, B. L. & Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. en. *Journal of Medicinal Chemistry* **60,** 7393–7409. ISSN: 0022-2623, 1520-4804. doi:10.1021/acs.jmedchem.7b00696 (Sept. 2017)

Chapter 3

Cáceres, E. L., Tudor, M. & Cheng, A. C. Deep learning approaches in predicting ADMET properties. *Future Medicinal Chemistry* **12.** PMID: 33124448, 1995–1999. doi:10.4155/fmc-2020-0259 (2020)

Chapter 4

Cáceres, E. L., Kuo, S. Y., Gestwicki, J. & Keiser, M. J. Target deconvolution for reduction offree tau in a high throughput screen. *Unpublished*

Epigraph

TU DIJISTE que amabas a la alondra por sobre todos los pájaros, por su vuelo recto hacia el sol. Así querías que fuese nuestro vuelo.

[...]

Sólo al morir tenemos aquel vuelo; ya el cuerpo no se apega nunca más a nosotros como tierra pesada de surco.

La Alondra, Gabriela Mistral

Methods for Cheminformatic Prediction

Elena Louise Cáceres

Abstract

Large scale biological datasets are often comprised of observations which are noisy, which are biased by environment or process, and which represent fragments of a perpetually growing, yet incomplete record of human knowledge. Changes to computational methods, data deposition and storage, and improved collection of data have the potential to mitigate some of these problems. However, no one solution works for all problems, and care must be taken to ensure that a chosen method to make predictions for small molecules will be effective.

This thesis centers itself on prediction. How do we improve screening predictions made on biased and incomplete information? How do we better represent fingerprints for ligand-based screening when molecular shape is important? What methods might best inform our ability to make predictions now and improve the next step in the future? And, how can we create testable hypotheses from phenotypic observations when we can't directly observe the mechanism of action?

Chapter 1 presents the published work "Adding Stochastic Negative Examples into Machine Learning Improves Molecular Bioactivity Prediction". To address concerns over the effect that biased molecular affinity datasets may have on the accuracy of deep learning models, this work suggests an online method where to improve prediction when a dataset is made up of more binders than non-binders. The method, SNA, samples random, unannotated compounds and assigns them as non-binders during neural network training. SNA drastically improves the ability of the network to identify false positives in a full matrix of drugs and protein binders while slightly hurting performance on a temporal split.

Chapter 2 encompasses published work "A Simple Representation of Three-Dimensional

Molecular Structure" which presents Extended Three-Dimensional Fingerprint (E3FP). This molecular fingerprinting technique generates a fingerprint that can represent three-dimensional structure for statistical and machine learning methods. Its advantages over two-dimensional fingerprints include the ability to encode structural relationships within a molecule and aggregation of fingerprints into a molecular ensembles. The E3FP was compared against existing two- and three-dimensional representations, and Chapter 2 shows some cases where the method outperformed these existing techniques.

Chapter 3 provides a brief commentary on the current outlook of deep learning for prediction of Adsorption, Distribution, Excretion, Metabolism, and Toxicity (ADMET). It describes how changes to molecular representations of molecules for deep learning have improved prediction of ADMET endpoints. It speculates on why techniques like neural network multitask training may be fall short of expectations when implemented in practice. And, it pushes for deep learning interpretability and error estimation to improve trust in deep learning models and to facilitate iterative improvement of models.

Finally, unpublished work in Chapter 4 focuses on how to use high throughput screening data to predict and rank a set of proteins to describe the clearance of free tau associated with applications for Alzheimer's Disease.

Contents

List of Figures

List of Tables

Chapter 1

Adding Stochastic Negative Examples into Machine Learning Improves Molecular Bioactivity Prediction

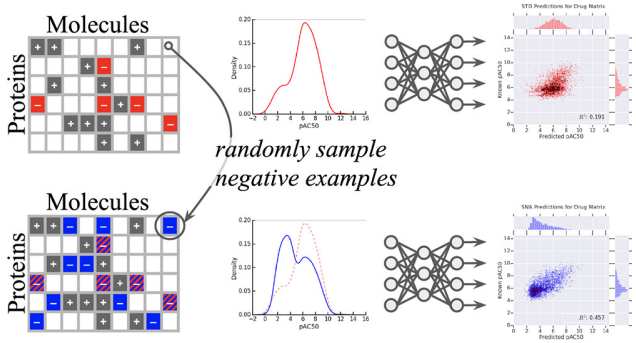

Figure 1.1: Stochastic Negative Addition of randomly-assigned negative examples changes the distribution of affinities in the minibatch, resulting in improved DNN predictive performance on a Drug Matrix hold out.

1.1 Abstract

Multitask deep neural networks learn to predict ligand–target binding by example, yet public pharmacological data sets are sparse, imbalanced, and approximate. We constructed two hold-out benchmarks to approximate temporal and drug-screening test scenarios, whose characteristics differ from a random split of conventional training data sets. We developed a pharmacological data set augmentation procedure, Stochastic Negative Addition (SNA), which randomly assigns untested molecule–target pairs as transient negative examples during training. Under the SNA procedure, drug-screening benchmark performance increases from $R^2 = 0.1926 \pm 0.0186$ to 0.4269 ± 0.0272 (122%). This gain was accompanied by a modest decrease in the temporal benchmark (13%). SNA increases in drug-screening performance were consistent for classification and regression tasks and outperformed y-randomized controls. Our results highlight where data and feature uncertainty may be problematic and how leveraging uncertainty into training improves predictions of drug–target relationships.

1.2 Introduction

Machine learning and deep neural network (DNN) methods have made great strides in scientific pattern recognition, particularly for cheminformatics[1–7]. As larger amounts of training data (molecules and their protein binding partners) have become publicly available, ligand-based predictions of polypharmacology have expanded from classification of binding (e.g., active/inactive) to regression of drug–target affinity scores (e.g., K_i and IC_{50})[3,4,8–12] These models exploit the similar property principle of chemical informatics, which states that small molecules with similar structures are likely to exhibit similar biological properties, such as their binding to protein targets[13]. Such approaches assume that the principle holds true for large data sets and hinge on the expectation that a greater diversity of training examples will increase the likelihood of a model finding generalizable patterns relating chemical structure

to bioactivity. However, these techniques may learn biased patterns from incomplete data for drug discovery and screening[14].

Bias for *in silico* quantitative structure–activity relationship (QSAR) data sets may be derived from a variety of sources including the publication record and scaffold bias. In industry, researchers frequently have access to large databases with large numbers of high-throughput screen examples that overwhelmingly comprise negative data. While these data span a diverse chemical space, researchers may be concerned about a bias toward previously studied scaffolds and programs. On the other hand, academic cheminformatic machine learning training sets may be typically derived from smaller, institutional data sets and sparse public bioactivity databases such as ChEMBL and PubChem BioAssay (PCBA)[15,16]. Theoretically, the more the researchers who deposit their data into these repositories, the more diverse the database. However, as scientific literature is a major contributor to these databases, any publication bias toward well-studied molecules or those with positive binding profiles (**Fig. 1.2**) skews both the data set and, consequently, the resulting machine learning models predictions, as reported by Kurczab et al[17]. We explore the feasibility of a method that leverages uncertainty in unexplored chemical space to augment incomplete public data for small molecule activity prediction using deep learning for both classification and regression.

A substantial literature focuses on correcting the balance of positive to negative examples (here, binders to nonbinders) in machine learning training data sets and addressing data set sparsity[12,18–25]. These corrections primarily adopt majority- or minority-based approaches. Minority-based approaches oversample underrepresented classes, and are generally accomplished by upweighting or oversampling existing training examples – or by adding similar synthetically generated ones[19,24]. Majority-based approaches typically undersample the overrepresented class in order to achieve balance. Many class imbalance approaches address situations where positive examples are in the minority. This presents a unique problem for cheminformatic data sets where binders ($<$10 μM) are frequently the majority class and non-

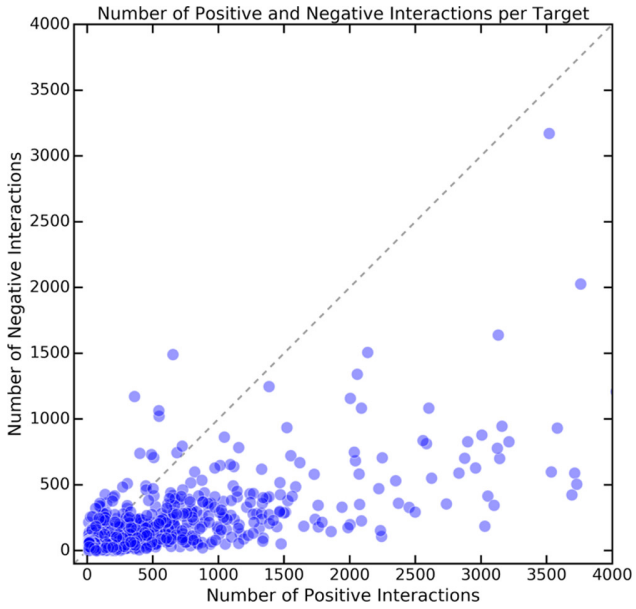

Figure 1.2: Protein targets are biased for positive interactions in a ChEMBL20. Target-wise distribution of binding (ligand) vs nonbinding molecules in the training set. Each point represents a single protein target drawn from ChEMBL20. A 1:1 ratio of binders (positives) and nonbinders (negatives) would fall along the dotted line.

binders are the minority reported class (**Fig. 1.2**), despite binding in comprehensive screens is a rare event[26]. For cheminformatic data sets, undersampling the majority class members could minimize the crucial effort researchers have invested to establish the chemical feature diversity upon which similar property principle-based approaches rely[25]. Accordingly, de la Vega de León et al. found that removing/ignoring activity labels can decrease performance in proportion to the amount of data removed[22]. As nonbinding molecules typically arise from the same series as binders, and consequently share many of their chemical features, we suspected that oversampling existing negative examples would contribute little to the expansion of a model's decision boundary. It follows, therefore, that oversampling may fail to add diversity to the minority class, whereas methods that rely on synthetic interpolation (i.e., generating new fingerprints very similar to existing negatives) increase the chance of mislabeling a new ligand in the chemical series and overlook protein targets lacking negative pharmacology data[19]. From a machine learning perspective, this hinders a model's generalizability and the scope of its chemical feature space, so oversampling negative examples

would seem especially problematic for cheminformatic data sets.

Random sampling of unassayed chemical space to assert weak but diverse negative examples may address these concerns. Others have shown that incorporating random negative data into training improves classification performance by SVMs[27], potency-sensitive influence relevance voters[28], and Bernoulli Näive Bayes classifiers[29]. Kurczab and Bojarski assessed the influence of negative data on a set of eight targets and found that a ratio of 9:1 to 10:1 of negatives to positives was favorable for classification[30]. In this work, we introduce putative negatives that continuously change throughout training and extend this method beyond classification to regression tasks for thousands of protein targets at once. We evaluate prediction performance on screening and temporal benchmarks and search for optimal positive-to-negative ratios under both test scenarios.

We propose an online (continuous) pharmacological training augmentation procedure for regression and classification tasks: stochastically oversampling the minority (nonbinder) class from the pool of unlabeled molecule-to-protein interactions spanning the molecule versus the protein target training space. We designed Stochastic Negative Addition (SNA) with the challenges of ligand-based drug design in mind. SNA adds more molecule–protein pairs to a training set where negative examples are otherwise outnumbered and/or unevenly distributed. Paradoxically, whereas most molecules do not bind to most proteins, the literature-based pharmacological data sets we used contain a preponderance of positive reports (**Fig. 1.2**); we intended SNA to counter this trend without overwhelming training with negative examples. This method encodes uncertainty for unstudied, and unlabeled, drug–target pairs. It exploits the observation that, despite meaningful cases of unexpected polypharmacology, ligand binding events at $\leq$10 μMare comparatively rare[26]. This study expands on prior work by investigating the effect of training augmentation for large numbers of protein targets in a multitask setting, applying the method to regression tasks and assessing the impact of random negatives on complementary benchmarks.

We assessed DNN model performance on two external test sets. We created a Time Split hold-out to address a drug discovery scenario with the understanding that this test set would be skewed toward having fewer negatives. However, while the Time Split might model drug discovery well, it has an unrealistic class balance (**Table A.4**). Therefore, we also created a complementary "screening" use-case benchmark, with a preponderance of negatives. We used the densely assayed Drug Matrix collection[31,32] and removed all of its protein–molecule interactions from the training set to avoid data leakage. We evaluated the Drug Matrix hold-out in tandem with Time Split for its ability to model screening cases where a researcher might wish to understand molecular binding profiles across a range of targets and compound libraries, where no prior publication reporting biases the benchmark in favor of binders. To determine how much pre-existing negative examples contributed to performance, we trained alternative DNNs where we removed negatives from the training data set. We explored whether SNA could rescue performance in this scenario where actual negatives were absent. We compared these models to an unaugmented, standard training regime and appropriate adversarial control studies[33]. We then explored whether different ratios of binders to nonbinders their affected performance. Finally, we evaluate whether SNA improves classification to the extent that it leads to regression. We find using SNA with a one-to-one positive-to-negative ratio improves performance on screening scenarios with minor penalty to temporal benchmarks.

1.3 Results

1.3.1 Adding Stochastic Negatives Improves Regression Performance

We posited that existing sparse public data sets omit much of the chemical diversity of the negative bioactivity space. To address this, we developed a machine learning training

procedure to transiently add likely negative examples: unstudied pairs of small molecules and protein targets that we assert to not bind. Using a SNA procedure, model predictions on a screening scenario benchmark data set (Drug Matrix) improved with minimal loss to performance on a temporal test benchmark (Time Split) (**Fig. 1.3**).

DNN models trained with five-fold cross validation using SNA (hereafter denoted in italics as SNA) outperformed conventionally trained models (standard; STD) on the screening (Drug Matrix) benchmark (**Fig. 1.3**e,f, **Tables 1.1** and **A.1** and **Figs. A.9**, **A.10**, **A.17** and **A.18**) with little effect on training or random validation performance (**Fig. 1.3**a–d, **Tables 1.1** and **A.1** and **Figs. A.9**, **A.10**, **A.17** and **A.18**). SNA performance increased by 122% in R^2 over the STD model on Drug Matrix affinity pAC_{50} values (see Methods). As with most screens, much of the data within the screening benchmark consisted of first-pass "primary" observations assessed only at a single dose of 10 μM. Regression could not be performed on these observations as no dose–response curve had been collected. To assess the effect of the proposed SNA training procedure on classification tasks, which would include these cases as well, we used two analyses: classification and regression-as-classification. The former consisted of training equivalent DNN architectures with classification loss functions—see the dedicated section below. For the latter, we evaluated the output of the original regression models as classifiers post hoc by thresholding affinity into positive and negative assignments according to pAC_{50} for the underlying truth values and constructing AUPRC_r and AUROC_r metrics over a range affinity thresholds instead of confidence thresholds. Thus, we calculated regression-as-classification AUPRC_r and AUROC_r by combining primary negatives with secondary (dose–response) negatives from the Drug Matrix screen versus secondary positives (see Methods). This analysis on Drug Matrix showed a 196% increase in AUPRC and 14% increase in AUROC_r for models trained using SNA over STD (**Fig. 1.3**i,k, **Tables 1.1** and **A.1** and **Figs. A.9**, **A.10**, **A.17** and **A.18**). A subsequent analysis incorporating SNA into a model trained on a scaffold split data set (see Butina Scaffold Split), showed results similar to Drug Matrix performance likewise from R^2 0.1547 $\pm$

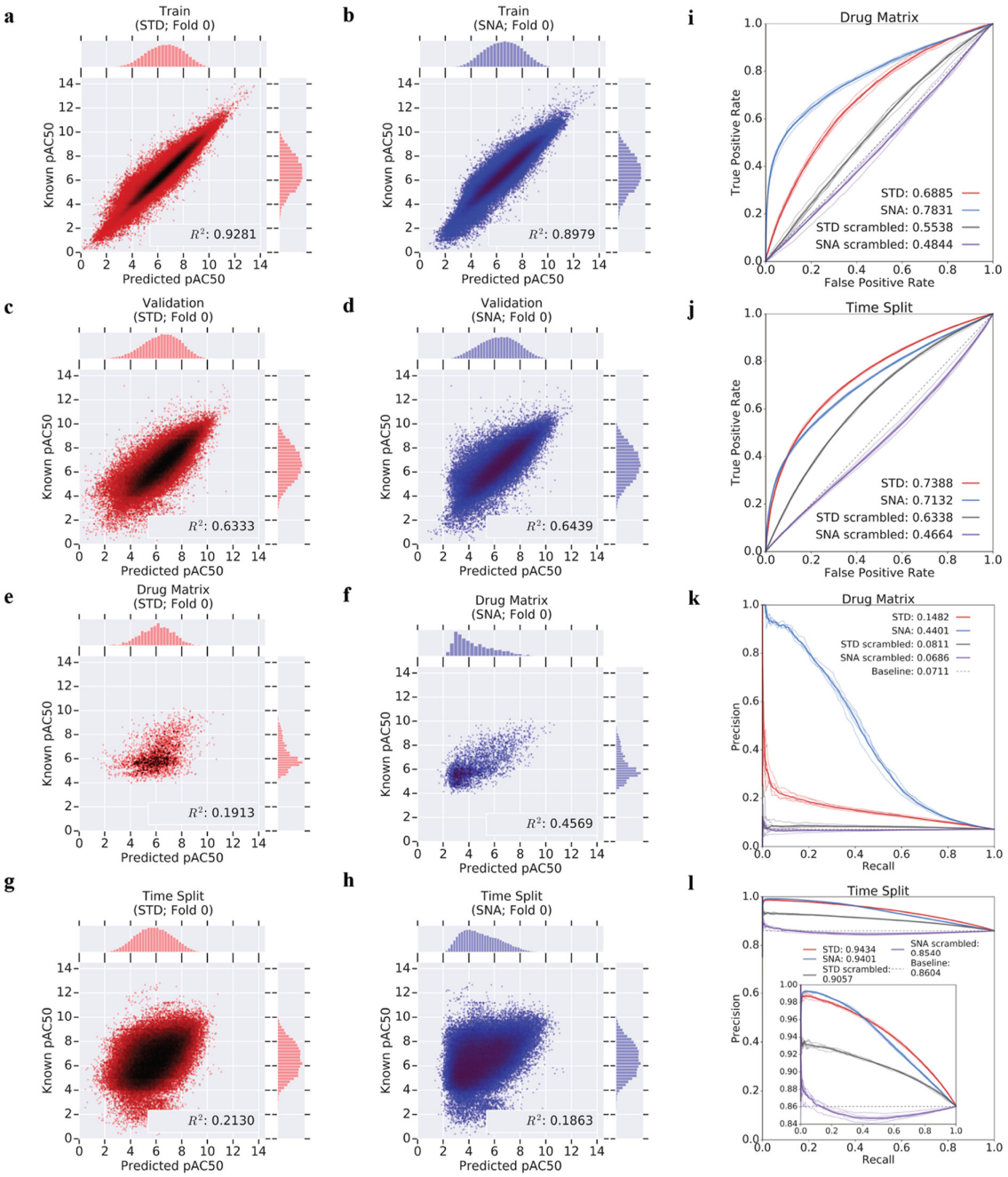

Figure 1.3: SNA markedly improves screening performance at minimal cost to temporal benchmarks. Predictions from a regression-based DNN model [STD; red; (a,c,e,g)] and the same model trained with the addition of stochastically chosen negative examples during each epoch [SNA; blue; (b,d,f,h)] show R^2 improvements on Drug Matrix (e,f) with minimal cost to Time Split (g,h) performance. Similarly, Drug Matrix screening benchmark by regression-as-classification $AUROC_r$ (i) and $AUPRC_r$ (k) plots, with Time Split $AUROC_r$ (j) and $AUPRC_r$ (l) plots favor SNA. Performance outperforms scrambled controls (STD scrambled—gray; SNA scrambled—purple), with SNA models showing greater gains over their random counterparts.

Table 1.1: Mean and Standard Deviation (std) Across Independent five-fold Cross Validation for STD and SNA DNN Models

Training Type	**Data Set**	**R^2 mean**	**R^2 std**	**$AUROC_r$ mean**	**$AUROC_r$ std**	**$AUPRC_r$ mean**	**$AUPRC_r$ std**
STD	Drug Matrix	0.1926	0.0186	0.6886	0.0094	0.1490	0.0077
STD scrambled		0.0154	0.0092	0.5538	0.0099	0.0816	0.0046
SNA		**0.4269**	**0.0272**	**0.7833**	**0.0059**	**0.4405**	**0.0079**
SNA scrambled		0.0021	0.0023	0.4842	0.0134	0.0687	0.0030
STD	Time Split	**0.2152**	**0.0033**	**0.7388**	**0.0024**	**0.9434**	**0.0008**
STD scrambled		0.0513	0.0032	0.6340	0.0033	0.9057	0.0010
SNA		0.1863	0.0012	0.7133	0.0025	0.9401	0.0006
SNA scrambled		0.0020	0.0016	0.4664	0.0106	0.8540	0.0032
STD	Validation	**0.6370**	**0.0041**	**0.9036**	**0.0016**	0.9837	0.0004
STD scrambled		0.0741	0.0026	0.6584	0.0014	0.9200	0.0019
SNA		**0.6428**	**0.0058**	**0.9064**	**0.0026**	**0.9848**	**0.0003**
SNA scrambled		0.0009	0.0004	0.4700	0.0053	0.8685	0.0012
STD	Train	**0.9224**	**0.0095**	**0.9809**	**0.0026**	**0.9972**	**0.0004**
STD scrambled		**0.9212**	**0.0016**	**0.9814**	**0.0002**	**0.9973**	**0.0000**
SNA		0.8971	0.0100	0.9750	0.0025	0.9962	0.0004
SNA scrambled		0.8618	0.0217	0.9725	0.0047	0.9958	0.0007

0.0026 to 0.3939 ± 0.0521 (154% increase) (**Table A.7** and **Figs. A.25** and **A.35** to **A.38**).

By contrast, SNA performance on the temporal (Time Split) benchmark decreased slightly, with SNA models decreasing by 13% in R^2 and 3% in $AUROC_r$ compared to STD (**Fig. 1.3**g,h,l, **Tables 1.1** and **A.1** and **Figs. A.9**, **A.10** and **A.17**). STD and SNA models generalized similarly on cross-validation sets (**Fig. 1.3**c,d, **Tables 1.1** and **A.1** and **Figs. A.9**, **A.10**, **A.17** and **A.18**), whereas standard models more precisely recapitulated their exact training examples [(**Fig. 1.3**a,b), **Tables 1.1** and **A.1** and **Figs. A.9**, **A.10**, **A.17** and **A.18**] than the equivalent SNA model, as expected. As Time Split chemical diversity may not directly reflect that of an explicit chemical scaffold split procedure, we subsequently created a Taylor–Butina clustered hold-out test set from the original data set (excluding Drug Matrix) and found trends similar to Time Split, albeit with higher overall performance, despite a 13% drop in R^2 from 0.3692 ± 0.0027 with STD to 0.3201 ± 0.0021 with SNA (**Table A.7** and **Figs. A.29** and **A.35** to **A.38**). We saw the same for the accompanying Taylor–Butina trained models, where SNA k-fold cross-validation performance for R2 was approximately 15% lower than STD R^2 of 0.3229 ± 0.0111 (**Table A.7**

and **Figs. A.30** and **A.35** to **A.38**).

SNA Brings Scrambled Control Models Closer to Theoretical Random for Regression

To evaluate whether the models withstood adversarial controls[33], we trained models on molecules whose annotations to protein targets had been randomized (y-randomization)[34–36]. SNA scrambled and STD scrambled models were trained with and without SNA procedures, respectively. Our goal was to verify that these intentionally scrambled models would underperform equivalent non-scrambled models on actual benchmarks. Thus, as in previous sections, we evaluated these models on screening, temporal, and fivefold cross-validation (Validation) sets.

As intended, scrambled models greatly underperformed those trained on data that was not scrambled (**Figs. 1.4**, **A.17** and **A.18** and **Tables 1.1** and **A.1**). However, some empirically scrambled models using standard training exceeded expected theoretical performance for balanced models (**Fig. 1.4**e–h; **Tables 1.1** and **A.1** and **Figs. A.17** and **A.18**). Scrambled models converged during training and achieved high performance on their scrambled train data sets (**Table A.1** and **Figs. A.13**, **A.14**, **A.17** and **A.18**), consistent with potential data set memorization rather than generalization[37]. Unsurprisingly, the R^2 for scrambled models neared 0.0 for screening, temporal, and cross-validation sets (**Figs. A.13** to **A.16**). While models trained on data that was not scrambled data outperform their scrambled controls, these controls exceeded frequently used, theoretical baselines such as 0.5 for $AUROC_r$ and the positive-to-negative ratio random baseline for $AUPRC_r$ in regression-as-classification analyses. STD scrambled models outperformed the 0.5 theoretical-random in $AUROC_r$ (Drug Matrix screening set: 0.5538 ± 0.0099; Time Split temporal set: 0.6340 ± 0.0033) (**Fig. 1.4**e,f; **Table A.1**). We also found that these STD scrambled models performed better than the random prevalence line in $AUPRC_r$ for Drug Matrix (random

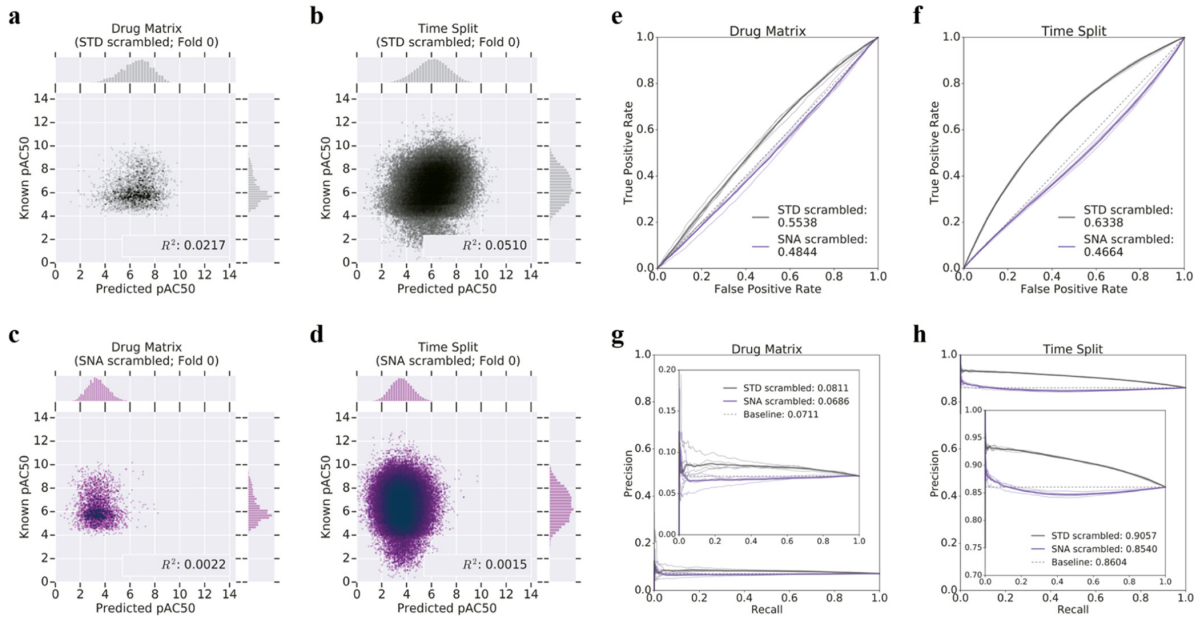

Figure 1.4: Scrambled control models with SNA more closely match expected random baselines. R^2 plots of a representative fold from fivefold cross validation (a–d) assess conventional DNN models trained with y-randomized data [STD scrambled; gray; (a,b)] and the equivalent networks trained with the stochastic negative procedure [SNA scrambled; purple; (c,d)]. Baselines for regression-as-classification $AUROC_r$ (e,f) compare the SNA scrambled random line (purple) to STD scrambled (gray) for both benchmarks. Regression-as-classification for $AUPRC_r$ (g,h) plots show a similar trend with respect to the ratio of positives-to-negatives in each benchmark.

prevalence: 0.0711; $AUPRC_r$: 0.0816 ± 0.0046) and temporal benchmarks (random prevalence: 0.8604; $AUPRC_r$: 0.9057 ± 0.0010) (**Fig. 1.4**g,h; **Table A.1**). SNA scrambled models exhibited reported values nearer the random baselines of 0.5 for $AUROC_r$ and the positive-to-negative ratio for $AUPRC_r$ in both the Drug Matrix benchmark ($AUROC_r$: 0.4842 ± 0.0134, $AUPRC_r$: 0.0687 ± 0.003) and temporal benchmark ($AUROC_r$: 0.4664 ± 0.0106, $AUPRC_r$: 0.8540 ± 0.0032) (**Fig. 1.4** and **Table A.1**).

For DNNs trained with Taylor–Butina k-fold cross-validation data sets, SNA moved the scrambled baseline toward 0.5 for $AUROC_r$ and to the positive-to-negative ratio for $AUPRC_r$, although this trend was less pronounced than in the random split cross-validation, particularly for Drug Matrix (**Table A.7** and **Figs. A.28** and **A.35** to **A.38**).

1.3.2 SNA Improves Performance for Classification Models

To evaluate whether the SNA training procedure was stable beyond regression and regression-as-classification, we developed and evaluated DNN classifiers with similar architectures. As with regression models, SNA classifiers saw increased model performance for the Drug Matrix screening benchmark, with a minor decline in the Time Split temporal benchmark (**Table A.2** and **Figs. 1.5**, **A.17** and **A.18**). In fivefold cross-validation, SNA improved screening benchmark performance by 151% AUPRC and 13% AUROC (**Table A.2** and **Figs. A.17** and **A.18**). Consistent with regression models, classification networks trained with SNA exhibited minor (4% AUROC and a 1% AUPRC) decreases on the Time Split benchmark (**Table A.2** and **Figs. A.17** and **A.18**). As before, both models outperformed their scrambled baselines. Classifier DNNs showed less performance gain over random controls in the temporal benchmark than regressor DNNs (**Fig. 1.5**e,f).

1.3.3 SNA Improves Regression Models Trained without Negatives

As SNA improved performance on a training set – where negatives are not guaranteed to be distributed across the benchmark sets in the same manner as the train set – we were curious whether SNA would improve cases where there are no true negative training data for ligand-binding prediction. To address this question, we evaluated two training regimes. First, we trained a DNN model solely on positive ligand–target examples (Negatives Removed). Second, we trained the equivalent Negatives Removed model, corrected by the SNA procedure (Negatives Removed + SNA). To compare model performance, we maintained the same benchmarks as before (screening/Drug Matrix, temporal/Time Split, and cross-validation). We hypothesized that removing all training negatives would damage model performance across the board, while incorporating SNA might partially rescue this effect. Additionally, we hypothesized classification models would be more sensitive to the removal

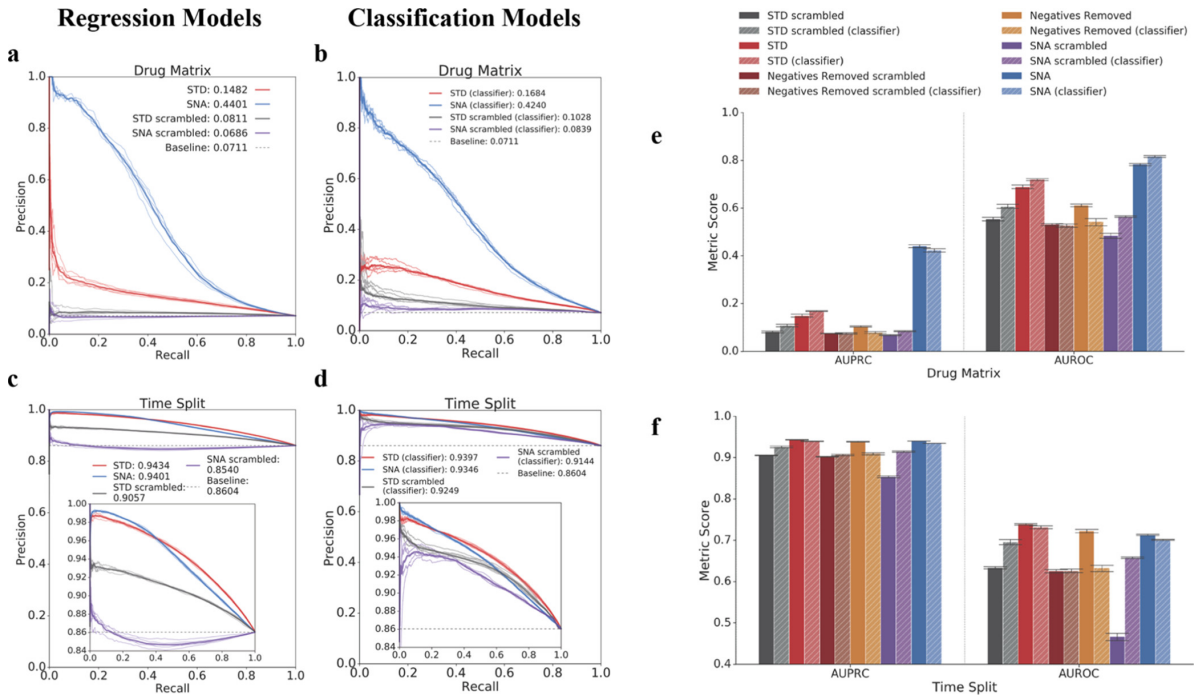

Figure 1.5: Classification models show similar trends to regression-as-classification evaluation. PRC on fivefold cross-validated classification networks (b,d,e,f) versus equivalent regression networks (a,c,e,f). For both classification (a,e) and regression tasks (b,e), networks trained with SNA (SNA; blue) achieved the highest Drug Matrix AUPRC. However, SNA models [blue; solid—regression, hashed—classification; (c,d)] did not show improved Time Split performance. AUROC and AUPRC bar plots for Drug Matrix (e) and Time Split (f) illustrate differences between classification (lighter, hashed bars) and regression models (solid bars) for SNA models (blue, green), their equivalent networks without SNA (red, orange), and y-randomized controls (grey, purple, sienna). All DNN models but the Negatives Removed classification model [(e,f); light orange; hashed] outperformed the scrambled benchmarks [(e,f); sienna; hashed]. Regression STD models (red; solid) underperformed classification STD models [light red; hashed; (e,f)], but the opposite was true for Negatives Removed models (orange; solid).

of training negatives than regression models.

Broadly, regression models trained without negative examples underperformed by regression-as-classification metrics, while achieving similar or better R^2 to standard (STD) for Drug Matrix and Time Split (**Fig. 1.6**; **Tables A.1** and **A.2**). The R^2 difference between Negatives Removed and STD models was minimal for the Drug Matrix screening benchmark (Negatives Removed R^2: 0.1973 ± 0.0176; STD R^2: 0.1926 ± 0.0186) (**Fig. 1.6**a; **Fig. 1.3**e; **Table A.1** and **Figs. A.9** and **A.11**). However, we observed larger differences in $AUROC_r$ and $AUPRC_r$, where the STD model outperformed the equivalent Negatives Removed model for Drug Matrix (Negatives Removed $AUROC_r$: 0.6120 ± 0.0076 vs STD $AUROC_r$: 0.6886 ± 0.0094; Negatives Removed $AUPRC_r$: 0.1039 ± 0.0025 vs STD $AUPRC_r$: 0.1490 ± 0.0077) (**Fig. 1.6**e, **Table A.1** and **Figs. A.23** and **A.24**). Removal of negatives from training harmed the Time Split temporal benchmark performance (−2.2% $AUROC_r$ and −0.5% $AUPRC_r$ change from STD) (**Fig. 1.4**f; **Table A.1**), but these models showed minor improvements in R^2 (9.3% increase from STD models) (**Fig. 1.6**c,f). For cross-validation (Validation) and training data benchmarks, removal of negatives during training uniformly decreased their performance by 5% (Validation) and 15% (Train) in R^2 (**Table A.1**).

We had anticipated that the SNA training procedure would partially mitigate the absence of true negatives during model training. Surprisingly, the Negatives Removed + SNA procedure yielded models with performance nearly indistinguishable from SNA models trained with full data, SNA (**Fig. 1.6**e,f, **Table A.1** and **Figs. A.1** to **A.4**). As with STD compared to SNA, Negatives Removed + SNA substantially improved the Drug Matrix screening benchmark performance while slightly decreasing that of the Time Split benchmark compared to a model trained with Negatives Removed alone. We observed 28, 331, and 116% increases to $AUROC_r$, $AUPRC_r$, and R^2, respectively, for the Drug Matrix screening benchmark by adding SNA training to Negatives Removed models (**Figs. 1.6** and **A.1** and **Table A.1**). By contrast, we observed that the Negatives Removed + SNA model training decreased temporal benchmark R^2 performance 25% to 0.1774 ± 0.0018 compared to the

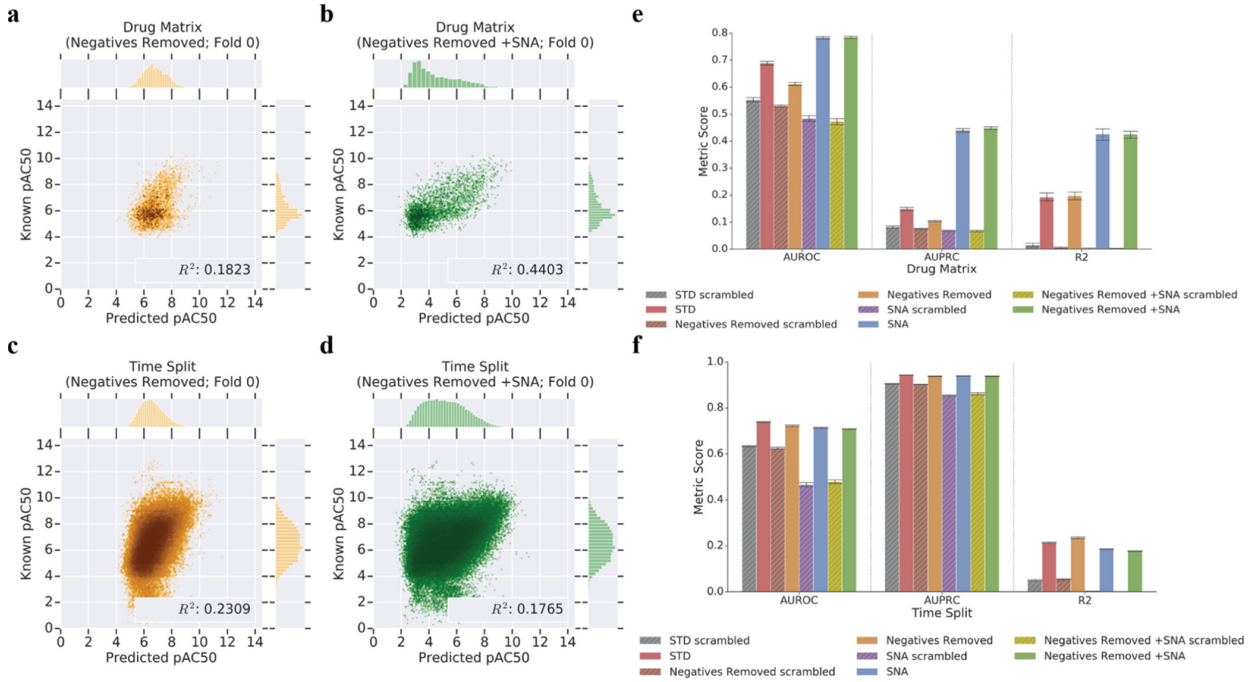

Figure 1.6: SNA rescues models trained without negative data. R^2 plots (a–e) of the same cross-validation fold show Negatives Removed + SNA models [green; (b,d)] rescue Drug Matrix performance, with a slight loss in Time Split performance. This is consistent across regression-as-classification $AUPRC_r$ and $AUROC_r$ metrics on fivefold cross validated networks (e). For Drug Matrix, SNA models [(e); blue, green] outperform equivalent conventional DNNs [(e); red, orange], but not for Time Split [(f); blue vs red, green vs orange]. Scrambled controls for each experiment [(e,f); gray, sienna, purple, chartreuse] establish baselines.

Negatives Removed alone (**Fig. 1.6**d,f; **Table A.1** and **Fig. A.2**). We found little to no change in mean AUC metrics for regression of Negatives Removed or Negatives Removed + SNA models, suggesting that neither stochastic nor true negatives improve performance on the temporal benchmark. Under these performance metrics, the impact of stochastic negative data on model training could not be distinguished from that of true negatives. However, we did not find that stochastic negatives yielded any greater performance than true negatives, despite the greater diversity of chemical examples covered by the former. To address whether there was an advantage from reported negatives, we performed an alternative training analysis wherein we upweighted existing negatives during training to reach parity between positives and negatives (see Negatives Upweighted; **Table A.1**) and found little improvement in the Time Split benchmark. To evaluate the influence of temporal versus scaffold test sets, we trained Negatives Removed and Negatives Removed + SNA models on a Taylor-Butina Scaffold Test Split data set. Again, we note large improvements to Drug Matrix R^2, $AUROC_r$, and $AUPRC_r$ (114%, 27%, and 292%, respectively) when Negatives Removed models are augmented with SNA (**Table A.7** and **Figs. A.28** and **A.35** to **A.38**). As with Time Split, Scaffold Split Test hold-outs performed worse or equivalently by R^2, $AUROC_r$, and $AUPRC_r$ (-24%, -0.1%, 0.5%) with SNA (**Table A.7** and **Figs. A.29** and **A.35** to **A.38**).

1.3.4 SNA Training Corrects for the Absence of True Negatives in Classification Nearly and in Regression

As with the regression models, removal of true negatives when training classification models affected performance in most benchmarks. SNA predominantly rescued performance for classification Negatives Removed models. The removal of true negatives from classification DNN training so adversely impacted performance on hold-out benchmarks that these models failed to exceed random baselines (**Tables A.2** and **A.8** and **Figs. A.21**, **A.22**, **A.33**

and **A.34**). This was consistent with the expectation that classification models trained solely on positive data would overwhelmingly predict positive outcomes. Therefore, we expected that incorporating stochastically imputed negatives during training (Negatives Removed + SNA) would improve classification. Drug Matrix screening benchmark performance improved markedly for Negatives Removed + SNA training compared to Negatives Removed models (48% increase in AUROC; 291% in AUPRC) (**Table A.2** and **Figs. A.21** and **A.22**). Negatives Removed + SNA only slightly improved Time Split AUROC and AUPRC (3% to AUROC and 1% to AUPRC), although this was in contrast to regression models, where SNA had decreased performance in this scenario (**Tables A.1** and **A.2** and **Figs. A.21** and **A.22**). Overall, we observed that the Negatives Removed regression model and its derived regression-as-classification interpretation outperformed the classification model on the screening benchmark. This was true also for Negatives Removed + SNA training with the exception of $AUROC_r$ for Drug Matrix.

1.3.5 Restricting SNA by Molecular Similarity Does Not Dramatically Improve the Procedure

To decrease the likelihood that SNA may assign true-but-unreported ligands to be negatives during training, we blocklisted potential molecule–target pairs by a separate cheminformatic method. This blocklist was created using the similarity ensemble approach (SEA) to predict likely binders (see SNA + SEAblocklisting). We assessed the networks trained with the SEA blocklist (SNA + SEA blocklist) similarly to the base SNA model procedure for both classification and regression. As with SNA networks, the SNA + SEA blocklist DNNs outperformed STD models on Drug Matrix with minor decreases to Time Split for regression (**Table A.1** and **Figs. A.1** and **A.2**) and classification (**Table A.2** and **Figs. A.5** and **A.6**). The performance differences between SNA and SNA + SEA blocklist were minimal, typically within a 1% difference (**Table A.1** and **Figs. A.1** to **A.4**). The exceptions

were $AUPRC_r$ and R^2, where SNA + SEA blocklist outperformed SNA on Drug Matrix by 2.8 and 3.3%, respectively. The same was true for classification networks, with SNA + SEA blocklist performing within a 1% difference to SNA for all but AUPRC, where SNA + SEA blocklist induced an increase in the mean across cross validated models of 3% (**Table A.2** and **Figs. A.5** to **A.8**).

1.3.6 Optimal SNA Ratio for DNN Performance Centers on 1:1 Positive/Negative Examples

To assess the impact of the class balance ratio chosen for the SNA training procedure, we trained 14 networks with SNA minimum ratios (i.e., minimum ratio of negatives-to-positives per protein target, below which negatives are added until the ratio is achieved) extending from no negatives added to the training (0% added) to 93%, as assessed on each protein target represented within a minibatch. We applied this procedure to regression and classification DNNs trained with STD and Negatives Removed contexts.

We found that the region between 40% and 60% added-negative ratio was the best tradeoff of performance across all benchmarks (**Fig. 1.7** and **Tables A.5** and **A.6**). Consistent with established class-balance training procedures, a 50% or 1:1 addition of SNA appears ideal, for both classification and regression scenarios. We note that the Drug Matrix screening benchmark improvement is the steepest between 10% and 30% negative addition; while the Time Split benchmark suffers some decreases in this regime, they are far less pronounced than the improvements to the screening benchmark.

The most exaggerated difference between classification and regression occurred for Negatives Removed models (**Fig. 1.7** and **Tables A.5** and **A.6**). Regression models trained using the Negatives Removed + SNA method almost entirely rescued the all-data SNA model performance by a 40% negative-addition ratio for Drug Matrix (**Fig. 1.7**b). However, classification Negatives Removed + SNA could not match the AUPRC Drug Matrix

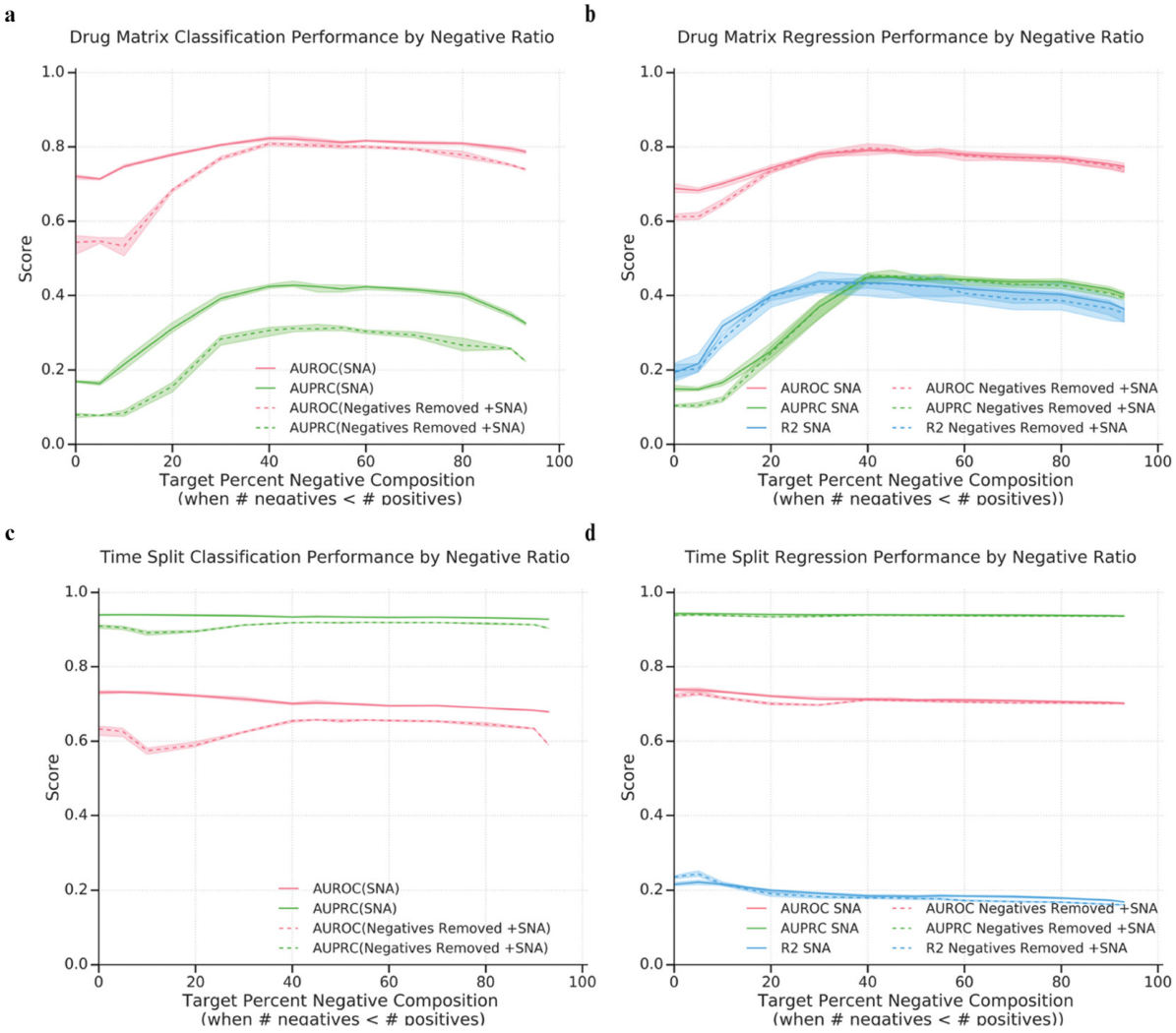

Figure 1.7: Balanced ratio of positives to negatives achieves the best overall performance. Increasing the targeted negative-to-positive ratio improves Drug Matrix performance for classification (a) and regression (b) up to 40–50% negatives per target, with modest impact to Time Split performance for classification (c) and regression (d) SNA models. SNA rescues the removal of reported negatives from model training data for regression (b,d) but not classification (a,c) (Negatives Removed + SNA; dashed line). Shaded areas represent the maximum and minimum boundaries within fivefold cross validation.

screening performance of STD (**Fig. 1.7**a). The screening benchmark performance difference between SNA and Negatives Removed + SNA models was less pronounced by AUROC. For the Time Split benchmark, stochastic negatives were less effective in closing the performance gap between Negatives Removed and STD classification models (**Fig. 1.7**c) than they were for regression models (**Fig. 1.7**d). From this, we conclude that classification tasks perform much better when trained on true negatives than when trained with stochastically imputed negatives. Regression models appear to see less gain from reported negatives as compared to stochastically imputed ones. However, as both classification and regression performed better with the addition of random negatives, we believe SNA can productively augment true negatives in the case when there are insufficient negatives and in the case when there are no negatives to speak of in the training set.

1.4 Discussion

We were concerned that insufficient proportions or diversity of negative examples in bioactivity training data sets could be lessening machine learning model precision for drug screening and discovery tasks. With this in mind, we hypothesized that adding transient random negative examples during training would improve model precision more than it would degrade model sensitivity. In this study, we set out to investigate the impact of SNA to DNN training for small-molecule-to-protein target affinities for classification and regression tasks. We found that adding stochastic negatives to DNN training improves predictive performance on a full-matrix screening benchmark (Drug Matrix; 7.1% positive cases, **Table A.4**) for both classification and regression tasks. This performance boost had minimal negative impact on a temporal evaluation scenario, which is skewed toward approximately 70% positive cases (**Table A.4**). We thresholded the regressed values analogously to classification networks to obtain a regression-based AUROC_r and AUPRC_r performance metrics, which frequently agree with classification performance trends. Finally, we compared the results to scrambled

baselines, which suggest the method does not solely rely on memorization for performance improvement.

The SNA data augmentation procedure improved both regression and classification DNN performance when compared to a standard model (STD) (**Figs. 1.3** and **1.5** and **Tables A.1** and **A.2**). Here, we define success as better performance on the screening-scenario benchmark, Drug Matrix, with a small relative hit to the Time Split benchmarks, which have a positive-to-negative ratio that favors positives. This suggests that SNA data augmentation allows a DNN to preserve potential QSAR information, yielding fewer false-positives that would otherwise plague the manual review of binding candidates. In a supplemental scaffold split analysis (see Butina Scaffold Split), we also found that the models perform better on scaffold-split testing than on Time Split, with similar minor performance decreases under SNA (**Tables A.1**, **A.2**, **A.7** and **A.8** and **Figs. A.2**, **A.6**, **A.17** to **A.21**, **A.23**, **A.24**, **A.26**, **A.29** and **A.31** to **A.38**). We found this behavior unexpected as prior analyses have shown temporal and scaffold splits to be similar in difficulty[5,38,39]. The use of Butina scaffold splitting to define the Validation set—and hence its convergence criteria—may contribute to the model's higher-than-expected performance in this scenario, as compared to Time Split, where the validation set was defined using a conventional random split. Under screening and scaffold split scenarios, the models meet anecdotal QSAR regression thresholds (R^2: 0.3) used for categorical analysis-based triage during early identification and prioritization of adsorption, metabolism, excretion, and toxicity profiles[40,41]. This capability is particularly useful for groups that do not have access to extensive and densely sampled corporate data sets. Performance by $AUROC_r$ and $AUPRC_r$ metrics indicate the gain to screening prediction capability for categorical or triaging decisions.

We observe that supplementing DNN training data with random negatives for a large number of targets from ChEMBL20 is consistent with previous findings with other machine learning models on smaller numbers of targets, which found that adding negative examples during training can improve classification of molecular bioactivity[17,30]. Despite the success of

minority upweighting in other machine learning fields, we find that upweighting negative data in the models' loss function during training yielded no advantage (**Table A.1** and **Figs. A.1** to **A.4**), suggesting that existing negatives alone do not cover either a sufficient number of targets or chemical feature space to be useful. This study adds to evidence, suggesting that balancing public data sets (given either positive or negative majority classes) for machine learning training improves classification prediction[27–29,42–44]. Here, we find a similar property holds for DNN regression, given a positive majority class, while including scrambled baselines to define random performance and show an added benefit of reduced memorization for augmented train sets.

While we find SNA improves Drug Matrix at a cost to Time Split performance, we note that the models contain negatives within the training set, which may be unevenly distributed across protein tasks. This distribution may artificially boost performance for certain tasks with additional negative data. To address this, we trained a model in the absence of negative data and without stochastic negatives as a sanity check, where we expected reduced performance due to loss of negative training examples for certain targets. We found that this Negatives Removed model improves generalizability for Time Split and depleted performance on Drug Matrix. We do not find this surprising, as the distribution of the Time Split data set more closely matches the Negatives Removed training set, possibly arising from a positive-reporting bias for novel molecules in the ChEMBL database underlying the temporal-split benchmark (71% positive, n = 116929, **Table A.4**), which is derived from the literature. A priori, a model trained without negative examples would be more likely to predict positive binding activity for a novel molecule; a benchmark with 71% positive cases would reward this propensity. As more negative and screening data make their way into public data sets, we will be interested to see whether this effect on temporal or scaffold splits will lessen over time.

Negatives Removed classification DNNs (**Table A.2** and **Figs. A.21** and **A.22**) performed far worse than Negatives Removed regression DNNs (**Fig. 1.6**e,f; **Table A.1**

and **Figs. A.23** and **A.24**). We hypothesize this may be because of different objectives between regression models and classification models—particularly in the absence of negative data. In such a scenario, removal of negative data is harmful to classification as there is no information from which to establish a binary decision boundary, whereas a regression network may be better equipped to extrapolate to unseen structures because of its modeling of continuous relationships between chemical structure and bioactivity.

Regression models trained with Negatives Removed exhibited performance losses which SNA rescued (**Fig. 1.6**). These data suggest that stochastic negatives may usefully supplement true negative data, but because of lack of clearly better performance, we do not believe that SNA should be used to supplant the use of true negatives in model training. SNA failed to completely rescue classification Negatives Removed performance. This may reflect fundamental differences between the aim of regression versus classification or the forms of the loss functions in question. Exploration of the ranked molecule choice between regression and classification models should be interesting for future *in silico* analyses. One explanation might be that underlying data set biases (such as molecular similarity) may have consequences for classification DNNs that are different for regression DNNs. Regardless, the data showing Negatives Removed + SNA rescuing model performance suggest it is reasonable to consider adding random negatives when none are available in the literature.

We also briefly explored the possibility that the SNA method of choosing potential negative pairs may have unintended consequences for ligands which are topologically similar to existing ligands for the same target. Using an alternative ligand–target prediction method, the SEA[8] to block potential molecule–target negative pairs and to reduce the probability of incorrectly assigning a likely ligand to be a negative example, we found that SNA + SEA blocklist models performed similarly to standard SNA equivalents (see Supplementary information for Chapter 1). From the results available, we see little reason to include SEA block-listing during training but see little reason to disavow exploration of additional blocklisting techniques in the future, such as sphere exclusion for dissimilar compound sam-

pling(29) as this may change with the data set, representation, model, or task. We note that SEA is a ligand-based approach and may not yield a sufficiently orthogonal blocklist, especially considering that our neural networks are trained on ligand topology. Future studies may address this concern by incorporating biophysical models as a blocklisting methodology, but we leave the exploration of negative choice open as an option for further studies.

We created scrambled DNN models (e.g., STD scrambled; SNA scrambled) to serve as low-performance adversarial baselines for our experiments and evaluate them against the same hold-out benchmarks. These baseline control studies yielded two key observations. First, as both STD and SNA outperformed their relative scrambled controls, the DNN models here do not rely solely on memorization for their performance. Second, as SNA decreased the baseline down toward 0.5 for AUROC/AUROC_r and to the positive-to-negative ratio for AUPRC/AUPRC_r, we found the SNA training procedure widens the predictive gap between actual and random models, suggesting additional benefits compared to STD when solely considering performance metrics.

While it is not unreasonable to assume an SNA ratio mimicking the underlying distribution of positives-to-negatives would produce the best result, we instead found that a balanced training set performed well in our exploration of different ratios (**Fig. 1.7** and **Tables A.5** and **A.6**). We observed that for SNA and Negatives Removed + SNA models, a data set comprising approximately 40–60% negatives per target maximized performance for Drug Matrix and Time Split. These results were consistent across regression and classification networks. By outperforming analogous models trained on scrambled data sets, we posit the models have learned beyond simple memorization, such as target (task) distributions. Considering that the bulk of the improvement on Drug Matrix occurs between 10% and 40% stochastic negatives, we hypothesize that future Time Splits with progressively more negatives may favor ratios approaching 50% as well. The optimal class balance we found here is inconsistent with studies using different types of models, which suggest positive to negative ratios around 1:9–1:10 for SMO, Random Forest, Ibk and J48 algorithms[17]. How-

ever, our results are consistent with their conclusion that optimal ratios may change with choice of algorithm and the properties of a given data set. For instance, our methods involve multiple rounds of random resampling throughout each epoch of DNN training, and we are performing a regression task.

This study is not without caveats. As noted in Methods, data from ChEMBL is biased by both the researcher and assay, and we have made several assumptions in aggregating data sets. We took aggregate values (median) for duplicated molecule–protein pairs to avoid over-sampling, particularly well-studied pairs. We made further bulk assumptions about our data set by asserting a single negative binding threshold ($pAC_{50} = 5.0$; 10 μM) when evaluating the performance, agnostic-to-protein target. For certain proteins, a hit weaker than 10 μM may be desirable for a researcher, and for other proteins, a hit stronger than 1 nM may be the minimum affinity necessary to describe a hit. It would be interesting to consider protein-wise hit thresholds for future AUROC_r and AUPRC_r regression-as-classification analyses. Our models are additionally limited by the representation of our data sets. We did not add any structural protein information. This limits the total variance we could expect to derive from such a data set, but we believe our method has uses where structural information is unavailable or where a phenotype-based readout is desired. Furthermore, our choice of the ECFP4 molecular feature representation[45] does not include information that could be obtained from 3D fingerprints or graph convolutional methods[1,46,47]. Finally, while our study defines success by the marked improvement on the screening Drug Matrix benchmark with minimal loss to Time and Scaffold Split Test benchmarks, we acknowledge that different test sets and measurement statistics are use-case specific choices that must be set by the researcher. For this reason, we publish performance across all benchmarks in terms of R^2, AUROC, and AUPRC, where it is appropriate.

This method is intended as an interim measure to supplement data sets, while quality in vitro negative data may be collected and reported by experimental researchers. It is not intended as a cure-all for the lack of negative data. It may be informative to more

finely evaluate under what conditions experimental negatives most effectively impact model predictions and where stochastically asserted ones are sufficient. Analysis for the particular protein target profiles that benefit under SNA conditions remain as an avenue for future studies. For example, although SNA and SNA + SEA blocklist models perform similarly, highly promiscuous targets may suffer under SNA and may suffer less under SNA + SEA blocklist models. Although this study was designed for data sets containing an affinity distribution bias, it is possible that stochastic injection of diverse compounds stratified by other properties could assist in other types of bias as well.

1.5 Conclusions

The SNA approach is a pharmacological data-augmentation procedure for DNNs designed to randomly assert untested negatives for public data sets where negative data are otherwise lacking. In each training epoch, new negatives are drawn to ensure that any particular negative choice does not heavily influence the model. We evaluated SNA at multiple ratios of positives to negatives and found that a ratio around 1:1 is optimal. We compared SNA training for both classification and regression networks trained on ChEMBL20. We found that, generally, SNA improved predictions on a held-out screening-like benchmark (Drug Matrix) with minimal effect on a 20% Time Split hold-out. Effectively, this resulted in a lower false-positive rate for the screening scenario. Our random selection of negative data involved minimal computational overhead. Supplementation of DNN training with stochastic negatives provides an interim augmentation measure for data sets lacking diverse negative data until more experimental data become publicly available.

1.6 Methods

1.6.1 Data Description

We filtered the ChEMBL20 database[15] by small molecule-target affinities with a binding type "B" and reported affinity values of type IC_{50}, EC_{50}, K_i, or K_d. Adapting the ontology from Visser et al., we treat all K_i, K_d, IC_{50}, EC_{50}, and related values equivalently and broadly refer to the resulting annotations as "activity concentration 50%" (AC_{50}) values[16]. We removed molecules with MW 800 Da and protein targets with fewer than 10 positive interactions. We addressed over-weighting of well-studied molecule-to-target pairs by taking the median across repeated target–molecule pairs. ChEMBL qualifies affinity using the "Relation" parameter that reports whether the true value is greater than, less than, or equal to the reported value. For all relations except "equals," we added random noise to the values to express uncertainty (**Table A.3**). We transformed all AC_{50} values by −log10 to arrive at pAC_{50} values for training, such that $pAC_{50} > 10$ (i.e., <0.1 nM) would be considered a strong binder and $pAC_{50} < 5$ (i.e. >10 μM) would be considered inactive. For classification tasks, we used pAC_{50} 5.0 to establish positive/active class identity.

Inputs are represented as a 4096-bit RDKit Morgan Fingerprint with a radius of 2[45]. Predicted values are the log transforms of affinity as described above (pAC_{50}) at 2038 protein targets for each molecule.

Our literature-derived annotations mined from ChEMBL skew the training set toward positive examples, with 73% representing binding affinities at 10 M or lower, 55% of 1 μM or lower, and 34% of 100 nM or lower. The remaining 27% of the training examples are explicit negatives—molecules that failed to inhibit the tested protein target by at least 50% at 10 μM. Six percent of targets (138 proteins) have zero reported nonbinders weaker than 10 μM. For training purposes, no targets have zero reported binders.

1.6.2 Data Splits

The evaluation benchmarks—which assess two distinct use cases—draw on Drug Matrix [CHEMBL1909046] and a 20% Time Split hold-out[38] and are excluded from the train and cross-validation sets. For the Time Split hold-out, we set aside approximately the final five years of ChEMBL activities, as assessed by the first reported publication date for a given interaction between the molecule and protein target (see code). Like ChEMBL, the Time Split hold-out is sparsely populated by negative data, but unlike a randomly split ChEMBL hold-out, it contains more unique structures. Drug Matrix is a data set produced by Iconix Pharmaceuticals that reports in vitro toxicology data for 870 chemicals across 132 protein targets[48]. Of these 132, we used the 84 targets that passed filtering steps defined in Data Description in our training set from Drug Matrix as a way to measure how we perform on a set containing a higher ratio of negative data to positive data. Descriptions of positive and negative attributes for each split are available in **Table A.4**.

1.6.3 Stochastic Negative Addition

SNA for multitask DNN training is added in an online fashion where new negative training examples for molecule–protein pairs are generated at each epoch to achieve a desired ratio of positives to negatives for each target. For the baseline SNA model, negatives are selected randomly from all unlabeled pairs in the data set to fulfill the desired ratio of positives to negatives at the target of interest. To evaluate the impact of potentially misassigning hidden positive examples during training, we developed a second method using the SEA[8] to blocklist potential interactions during the sampling procedure (SNA + SEA blocklist). For this method, we excluded from consideration ("blocklisted") all otherwise unlabeled pairs that achieved a positive SEA prediction with pSEA 5. We tested SNA at the following positive-to-negative ratios to find an optimal balance beginning at the baseline positive prevalence in Drug Matrix: [0.07, 0.5, 0.6, 0.75, 0.85, 0.95, 1.0, 1.33, 1.54, 2.0, 2.86, 4.0,

6.66, 10.0].

1.6.4 Negatives Removed

To assess the impact of training on purely positive data, we scrubbed all negative data ($pAC_{50} < 5$) from the training set (Negatives Removed) and evaluated it on Time Split and Drug Matrix as we did for training regimes including negatives such as STD and SNA. We applied SNA to the training regime as above to evaluate the impact of stochastic negatives on the model's predictive ability (Negatives Removed + SNA). Each of the 14 ratios listed above were tested for SNA applied to models trained on the negatives-removed data set to evaluate the impact of positive-to-negative ratios on performance.

1.6.5 Software

This project was built with Python 2.7. All DNNs were implemented and trained in Lasagne[49] and Theano[50]. We used RDKit for all handling of molecular structures[51]. We used NumPy[52] and Scikit-learn[53] for performance measures and numerical analyses, and visualizations were made with Matplotlib[54] and Seaborn[55].

1.6.6 Multitask Deep Neural Network Model Hyperparameters and Architecture

As multitask DNN performance is sensitive to architecture and hyperparameter choice, we optimized hyperparameters and architecture by considering retrospective performance on a random 20% hold-out of the training data set. We performed a grid search over varying architectures and manually explored for optimal hyperparameters. Although this optimization is not exhaustive, we focused this study on a simple representative architecture with three fully connected hidden layers with 1024, 2048, and 3072 nodes, respectively. We used

an input layer of 4096 nodes, the length of our input fingerprints, and an output layer with 2038 target nodes. We use leaky rectified linear unit (leaky-ReLU) activation functions[56] for all hidden layers and L2 weight regularization with a penalty of 5 × 10–5 and mean squared error for the loss function. We employed stochastic gradient descent with Nesterov momentum[57] using a fixed learning rate of 0.01 and momentum of 0.4. Additionally, the hidden layers were subject to dropout[58] with probabilities of 0.1, 0.25, and 0.25, respectively.

1.6.7 Model Training and Classification Accuracy Assessment

R-Squared

We square the correlation coefficient (r_value) from scipy.stats.linregress.

Area under the Curve

Area under the curve (AUC) was analyzed for both the precision-recall curve (AUPRC) and the receiver operating characteristic curve (AUROC). For classification models, AUC was implemented as in sklearn, with a ground-truth positive threshold set to 5.0 as in training. While AUPRC and AUROC are traditionally reported for classification models, we also reported these metrics for thresholded regression models and denote these with AUROC_r and AUPRC_r for clarity. This usage has two underlying assumptions: (1) chemical screens are performed to assess hit rates past a certain biological threshold [e.g., p(AC_{50} in molar) 5] and (2) higher ranking predictions from a regression model are more likely to be tested first by researchers. Given these assumptions, we posthoc assessed regression models as classification. Prediction thresholds were chosen over the maximum and minimum of the predictions for a given model in step sizes of 0.05, and true values were thresholded at 5.0. At each prediction threshold, true positive rate/precision, false-positive rate, and recall were calculated and then the AUC generated from points.

1.6.8 Source Code

All code necessary to reproduce this work is available at `https://github.com/keiserlab/stochastic-negatives-paper` under the MIT License.

1.7 Supporting Information

The Supporting Information is available free of charge at `https://pubs.acs.org/doi/10.1021/acs.jcim.0c00565.`

1.8 Funding

This material is based on work supported by grant number 2018-191905 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation (M.J.K.) and a New Frontier Research Award from the Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation (M.J.K.). E.L.C. is supported under the National Science Foundation Graduate Research Fellowship Program under grant no. 1650113 and is a Howard Hughes Medical Institute Gilliam Fellow.

1.9 Acknowledgments

We thank Benjamin Wong, Brian Shoichet, Garrett Gaskins, Gregory Valiant, Jason Gestwicki, Kangway Chuang, Leo Gendelev, Jessica McKinley, and Michael Mysinger for discussion and technical support.

1.10 Abbreviations Used

1.10.1 General Abbreviations

SNA - Stochastic negative addition as a procedure

AUROC - AUC of the receiver operating characteristic curve

AUPRC - AUC of the precision-recall curve (classification)

$AUROC_r$ - AUC of the receiver operating characteristic curve (regression-as-classification)

$AUPRC_r$ - AUC of the precision-recall curve (regression-as-classification)

1.10.2 Model Abbreviations

STD - "standard" model trained without SNA procedure STD scrambled - STD model trained with y-randomization of the input training data

SNA scrambled - SNA model trained with y-randomization of the input training data

Negatives Removed - model trained with negatives removed from the training set

Negatives Removed scrambled - Negatives Removed model trained with y-randomization of the input training data

SNA + SEA blacklist - SNA model where ligands with a chance of binding (by SEA) are blacklisted from SNA choice during training

References

1. Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., Sun, S., Yang, J., Ramsundar, B. & Pande, V. S. PotentialNet for Molecular Property Prediction. en. *ACS Cent Sci* **4,** 1520–1530 (Nov. 2018) (cit. on pp. 2, 25).

2. Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A. & Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. en. *Chem. Sci.* **9,** 5441–5451 (June 2018) (cit. on p. 2).

3. Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D. & Pande, V. Massively Multitask Networks for Drug Discovery (Feb. 2015) (cit. on p. 2).

4. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. en. *J. Chem. Inf. Model.* **55,** 263–274 (Feb. 2015) (cit. on p. 2).

5. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. & Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. en. *J. Chem. Inf. Model.* **59,** 3370–3388 (Aug. 2019) (cit. on pp. 2, 21).

6. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackerman, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R. & Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. en. *Cell* **180,** 688–702.e13 (Feb. 2020) (cit. on p. 2).

7. Withnall, M., Lindelöf, E., Engkvist, O. & Chen, H. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *J. Cheminform.* **12,** 1 (Jan. 2020) (cit. on p. 2).

8. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. & Shoichet, B. K. Relating protein pharmacology by ligand chemistry. en. *Nat. Biotechnol.* **25,** 197–206 (Feb. 2007) (cit. on pp. 2, 28).

9. Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., Whaley, R., Glennon, R. a., Hert, J., Thomas, K. L. H., Edwards, D. D., Shoichet, B. K. & Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **462,** 175–181 (Nov. 2009) (cit. on p. 2).

10. Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Côté, S., Shoichet, B. K. & Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486,** 361–367 (June 2012) (cit. on p. 2).

11. Sydow, D., Burggraaff, L., Szengel, A., van Vlijmen, H. W. T., IJzerman, A. P., van Westen, G. J. P. & Volkamer, A. Advances and Challenges in Computational Target Prediction. en. *J. Chem. Inf. Model.* **59,** 1728–1742 (May 2019) (cit. on p. 2).

12. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. en. *Chem. Rev.* **119,** 10520–10594 (July 2019) (cit. on pp. 2, 3).

13. Johnson, M. A., Maggiora, G. M. & American Chemical Society. Meeting. *Concepts and applications of molecular similarity* en (Wiley, New York, 1990) (cit. on p. 2).

14. Ding, H., Takigawa, I., Mamitsuka, H. & Zhu, S. *Similarity-based machine learning methods for predicting drug–target interactions: a brief review* 2014 (cit. on p. 3).

15. Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R. & Overington, J. P. The ChEMBL bioactivity database: an update. en. *Nucleic Acids Res.* **42,** D1083–90 (Jan. 2014) (cit. on pp. 3, 27).

16. Visser, U., Abeyruwan, S., Vempati, U., Smith, R. P., Lemmon, V. & Schürer, S. C. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. en. *BMC Bioinformatics* **12,** 257 (June 2011) (cit. on pp. 3, 27).

17. Kurczab, R., Smusz, S. & Bojarski, A. J. The influence of negative training set size on machine learning-based virtual screening. en. *J. Cheminform.* **6,** 32 (June 2014) (cit. on pp. 3, 21, 24).

18. Japkowicz, N. & Stephen, S. The class imbalance problem: A systematic study. *Intelligent data analysis* **6,** 429–449 (2002) (cit. on p. 3).

19. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *1* **16,** 321–357 (June 2002) (cit. on pp. 3, 4).

20. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks (Oct. 2017) (cit. on p. 3).

21. Whitehead, T. M., Irwin, B. W. J., Hunt, P., Segall, M. D. & Conduit, G. J. Imputation of Assay Bioactivity Data Using Deep Learning. en. *J. Chem. Inf. Model.* **59,** 1197–1204 (Mar. 2019) (cit. on p. 3).

22. De la Vega de León, A., Chen, B. & Gillet, V. J. Effect of missing data on multitask prediction methods. en. *J. Cheminform.* **10,** 26 (May 2018) (cit. on pp. 3, 4).

23. Huang, C., Li, Y., Loy, C. C. & Tang, X. *Learning Deep Representation for Imbalanced Classification* in (2016), 5375–5384 (cit. on p. 3).

24. He, H. & Garcia, E. A. *Learning from Imbalanced Data* 2009 (cit. on p. 3).

25. Sundar, V. & Colwell, L. *The Effect of Debiasing Protein–Ligand Binding Data on Generalization* 2020 (cit. on pp. 3, 4).

26. Bradley, D. Dealing with a data dilemma. en. *Nat. Rev. Drug Discov.* **7,** 632–633 (Aug. 2008) (cit. on pp. 4, 5).

27. Heikamp, K. & Bajorath, J. Comparison of confirmed inactive and randomly selected compounds as negative training examples in support vector machine-based virtual screening. en. *J. Chem. Inf. Model.* **53,** 1595–1601 (July 2013) (cit. on pp. 5, 22).

28. Lusci, A., Browning, M., Fooshee, D., Swamidass, J. & Baldi, P. Accurate and efficient target prediction using a potency-sensitive influence-relevance voter. en. *J. Cheminform.* **7,** 63 (Dec. 2015) (cit. on pp. 5, 22).

29. Mervin, L. H., Afzal, A. M., Drakakis, G., Lewis, R., Engkvist, O. & Bender, A. Target prediction utilising negative bioactivity data covering large chemical space. en. *J. Cheminform.* **7,** 51 (Oct. 2015) (cit. on pp. 5, 22).

30. Kurczab, R. & Bojarski, A. J. The influence of the negative-positive ratio and screening database size on the performance of machine learning-based virtual screening. en. *PLoS One* **12,** e0175410 (Apr. 2017) (cit. on pp. 5, 21).

31. *DrugMatrix/ToxFX* `https://ntp.niehs.nih.gov/results/drugmatrix/index.html`. Accessed: 2019-6-23 (cit. on p. 6).

32. Svoboda, D. L., Saddler, T. & Auerbach, S. S. *An Overview of National Toxicology Program's Toxicogenomic Applications: DrugMatrix and ToxFX* 2019 (cit. on p. 6).

33. Chuang, K. V. & Keiser, M. J. Adversarial Controls for Scientific Machine Learning. en. *ACS Chem. Biol.* **13,** 2819–2821 (Oct. 2018) (cit. on pp. 6, 10).

34. Lipiński, P. F. J. & Szurmak, P. *SCRAMBLE'N'GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models* 2017 (cit. on p. 10).

35. Rücker, C., Rücker, G. & Meringer, M. y-Randomization and its variants in QSPR/QSAR. en. *J. Chem. Inf. Model.* **47,** 2345–2357 (Nov. 2007) (cit. on p. 10).

36. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. en. *Mol. Inform.* **29,** 476–488 (July 2010) (cit. on p. 10).

37. Wallach, I. & Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. en. *J. Chem. Inf. Model.* **58,** 916–932 (May 2018) (cit. on p. 10).

38. Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. en. *J. Chem. Inf. Model.* **53,** 783–790 (Apr. 2013) (cit. on pp. 21, 28).

39. Xu, Y., Ma, J., Liaw, A., Sheridan, R. P. & Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **57,** 2490–2504 (Oct. 2017) (cit. on p. 21).

40. Sanders, J. M., Beshore, D. C., Culberson, J. C., Fells, J. I., Imbriglio, J. E., Gunaydin, H., Haidle, A. M., Labroli, M., Mattioni, B. E., Sciammetta, N., *et al.* Informing the Selection of Screening Hit Series with in Silico Absorption, Distribution, Metabolism, Excretion, and Toxicity Profiles: Miniperspective. *J. Med. Chem.* **60,** 6771–6780 (2017) (cit. on p. 21).

41. Cáceres, E. L., Tudor, M. & Cheng, A. C. Deep learning approaches in predicting ADMET properties. en. *Future Med. Chem.* (Oct. 2020) (cit. on p. 21).

42. Korkmaz, S. *Deep Learning-Based Imbalanced Data Classification for Drug Discovery* 2020 (cit. on p. 22).

43. Zakharov, A. V., Peach, M. L., Sitzmann, M. & Nicklaus, M. C. QSAR modeling of imbalanced high-throughput screening data in PubChem. en. *J. Chem. Inf. Model.* **54,** 705–712 (Mar. 2014) (cit. on p. 22).

44. Lee, Y. O. & Kim, Y. J. The Effect of Resampling on Data-imbalanced Conditions for Prediction towards Nuclear Receptor Profiling Using Deep Learning. *Mol. Inform.* **39,** 1900131 (Aug. 2020) (cit. on p. 22).

45. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. en. *J. Chem. Inf. Model.* **50,** 742–754 (May 2010) (cit. on pp. 25, 27).

46. Axen, S. D., Huang, X.-P., Cáceres, E. L., Gendelev, L., Roth, B. L. & Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. en. *J. Med. Chem.* **60,** 7393–7409 (Aug. 2017) (cit. on p. 25).

47. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. en. *J. Comput. Aided Mol. Des.* **30,** 595–608 (Aug. 2016) (cit. on p. 25).

48. Ganter, B., Tugendreich, S., Pearson, C. I., Ayanoglu, E., Baumhueter, S., Bostian, K. A., Brady, L., Browne, L. J., Calvin, J. T., Day, G.-J., Breckenridge, N., Dunlea, S., Eynon, B. P., Furness, L. M., Ferng, J., Fielden, M. R., Fujimoto, S. Y., Gong, L., Hu, C., Idury, R., Judo, M. S. B., Kolaja, K. L., Lee, M. D., McSorley, C., Minor, J. M., Nair, R. V., Natsoulis, G., Nguyen, P., Nicholson, S. M., Pham, H., Roter, A. H., Sun, D., Tan, S., Thode, S., Tolley, A. M., Vladimirova, A., Yang, J., Zhou, Z. & Jarnagin, K. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. en. *J. Biotechnol.* **119,** 219–244 (Sept. 2005) (cit. on p. 28).

49. Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S. K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J. D., Heilman, M., de Almeida, D. M., McFee, B., Weideman, H., Takács, G., de Rivaz, P., Crall, J., Sanders, G., Rasul, K., Liu, C., French, G. & Degrave, J. *Lasagne: First release* Aug. 2015 (cit. on p. 29).

50. The Theano Development Team *et al.* Theano: A Python framework for fast computation of mathematical expressions (May 2016) (cit. on p. 29).

51. Landrum, G. *RDKit* 2010 (cit. on p. 29).

52. Oliphant, T. *Guide to NumPy: 2nd Edition* en (CreateSpace, 221 W. 6th Street, 15th Floor, Austin, TX, Sept. 2015) (cit. on p. 29).

53. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau,

D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011) (cit. on p. 29).

54. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9,** 90–95 (2007) (cit. on p. 29).

55. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A. & Qalieh, A. *mwaskom/seaborn: v0.8.1 (September 2017)* Sept. 2017 (cit. on p. 29).

56. Maas, A. L., Hannun, A. Y. & Ng, A. Y. *Rectifier nonlinearities improve neural network acoustic models* in *Proc. icml* **30** (2013), 3 (cit. on p. 30).

57. Nesterov, Y. *A method of solving a convex programming problem with convergence rate* $O\left(\frac{1}{k^2}\right)$ in *Soviet Math. Dokl* **27** () (cit. on p. 30).

58. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15,** 1929–1958 (2014) (cit. on p. 30).

Chapter 2

A simple representation of three-dimensional molecular structure

2.1 Abstract

Statistical and machine learning approaches predict drug-to-target relationships from 2D small-molecule topology patterns. One might expect 3D information to improve these calculations. Here we apply the logic of the Extended Connectivity FingerPrint (ECFP) to develop a rapid, alignment- invariant 3D representation of molecular conformers, the Extended Three-Dimensional FingerPrint (E3FP). By integrating E3FP with the Similarity Ensemble Approach (SEA), we achieve higher precision-recall performance relative to SEA with ECFP on ChEMBL20, and equivalent receiver operating characteristic performance. We identify classes of molecules for which E3FP is a better predictor of similarity in bioactivity than is ECFP. Finally, we report novel drug-to-target binding predictions inaccessible by 2D fingerprints and confirm three of them experimentally with ligand efficiencies from 0.442 - 0.637 kcal/mol/heavy atom.

2.2 Introduction

Many molecular representations have arisen since the early chemical informatics models of the 1970s, yet the most widely used still operate on the simple two-dimensional (topological) structures of small molecules. Fingerprints, which encode molecular 2D substructures as overlapping lists of patterns, were a first means to scan chemical databases for structural similarity using rapid bitwise logic on pairs of molecules. Pairs of molecules that are structurally similar, in turn, often share bioactivity properties[1] such as protein binding profiles. Whereas the prediction of biological targets for small molecules would seem to benefit from a more thorough treatment of a molecule's explicit ensemble of three-dimensional (3D) conformations[2], pragmatic considerations such as calculation cost, alignment invariance, and uncertainty in conformer prediction[3] nonetheless limit the use of 3D representations by large-scale similarity methods such as the Similarity Ensemble Approach (SEA)[4,5], wherein the count of pairwise molecular calculations reaches into the hundreds of billions. Furthermore, although 3D representations might be expected to outperform 2D ones, in practice, 2D representations nonetheless are in wider use and can match or outperform them[3,6–8].

The success of statistical and machine learning approaches building on 2D fingerprints reinforces the trend. Naive Bayes Classifiers (NB)[9–11], Random Forests (RF)[12,13], Support Vector Machines (SVM)[10,14,15], and Deep Neural Networks (DNN)[16–20] predict a molecule's target binding profile and other properties from the features encoded into its 2D fingerprint. SEA and methods building on it such as Optimized Cross Reactivity Estimation (OCEAN)[21] quantify and statistically aggregate patterns of molecular pairwise similarity to the same ends. Yet these approaches cannot readily be applied to the 3D molecular representations most commonly used. The Rapid Overlay of Chemical Structures (ROCS) method is an alternative to fingerprints that instead represents molecular shape on a conformer-by-conformer basis via gaussian functions centered on each atom. These functions may then be compared between a pair of conformers[22,23]. ROCS however must align conformers to de-

termine pairwise similarity; in addition to the computational cost of each alignment, which linear algebraic approximations such as SCISSORS[24] mitigate, the method provides no invariant fixed-length fingerprint (feature vectors) per molecule or per conformer for use in machine learning. One way around this limitation is to calculate an all-by-all conformer similarity matrix ahead of time, but this is untenable for large datasets such as ChEMBL[25] or the 70-million datapoint ExCAPE-DB[26], especially as the datasets continue to grow.

Feature Point Pharmacophores (FEPOPS), on the other hand, use k-means clustering to build a fuzzy representation of a conformer using a small number of clustered atomic feature points, which simplify shape and enable rapid comparison[27,28]. FEPOPS excels at scaffold hopping, and it can use charge distribution based pre-alignment to circumvent a pairwise alignment step. However, pre-alignment can introduce similarity artifacts, such that explicit pairwise shape-based or feature- point-based alignment may nonetheless be preferred[27]. Accordingly, 3D molecular representations and scoring methods typically align conformers on a pairwise basis[2,3]. An alternative approach is to encode conformers against 3- or 4-point pharmacophore keys that express up to 890,000 or 350 million discrete pharmacophores, respectively[29,30]. The count of purchasable molecules alone, much less their conformers, however, exceeds 200 million in databases such as ZINC (`zinc.docking.org`)[31], and the structural differences determining bioactivity may be subtle. To directly integrate 3D molecular representations with statistical and machine learning methods, we developed a 3D fingerprint that retains the advantages of 2D topological fingerprints. Inspired by the widely used circular ECFP (2D) fingerprint, we develop a spherical Extended 3D FingerPrint (E3FP) and assess its performance relative to ECFP for various systems pharmacology tasks. E3FP is an open-source fingerprint that encodes 3D information without the need for molecular alignment, scales linearly with 2D fingerprint pairwise comparisons in computation time, and is compatible with statistical and machine learning approaches that have already been developed for 2D fingerprints. We use it to elucidate regions of molecular similarity space that could not previously be explored. To demonstrate its utility, we combine E3FP

with SEA to predict novel target- drug activities that SEA could not discover using ECFP, and confirm experimentally that they are correct.

2.3 Results

The three-dimensional fingerprints we present are motivated by the widely-used two-dimensional (2D) Extended Connectivity FingerPrint (ECFP)[32], which is based on the Morgan algorithm[33]. ECFP is considered a 2D or "topological" approach because it encodes the internal graph connectivity of a molecule without explicitly accounting for 3D structural patterns the molecule may adopt in solution or during protein binding. While ECFP thus derives from the neighborhoods of atoms directly connected to each atom, a 3D fingerprint could incorporate neighborhoods of nearby atoms in 3D space, even if they are not directly bonded. We develop such an approach and call it an Extended Three-Dimensional FingerPrint (E3FP).

A single small molecule yields multiple 3D fingerprints

Many small molecules can adopt a number of energetically favorable 3D conformations, termed "conformers". In the absence of solved structures, it is not always apparent which conformer a molecule will adopt in solution, how this may change on protein binding, and which protein-ligand interactions may favor which conformers[34]. Accordingly, we generate separate E3FPs for each of multiple potential conformers per molecule. E3FP encodes all three-dimensional substructures from a single conformer into a bit vector, represented as a fixed-length sequence of 1s and 0s (**Fig. 2.1**a). This is analogous to the means by which ECFP represent two-dimensional substructures. To encode the three-dimensional environment of an atom, E3FP considers information pertaining not only to contiguously bound atoms, but also to nearby unbound atoms and to relative atom orientations (stere-

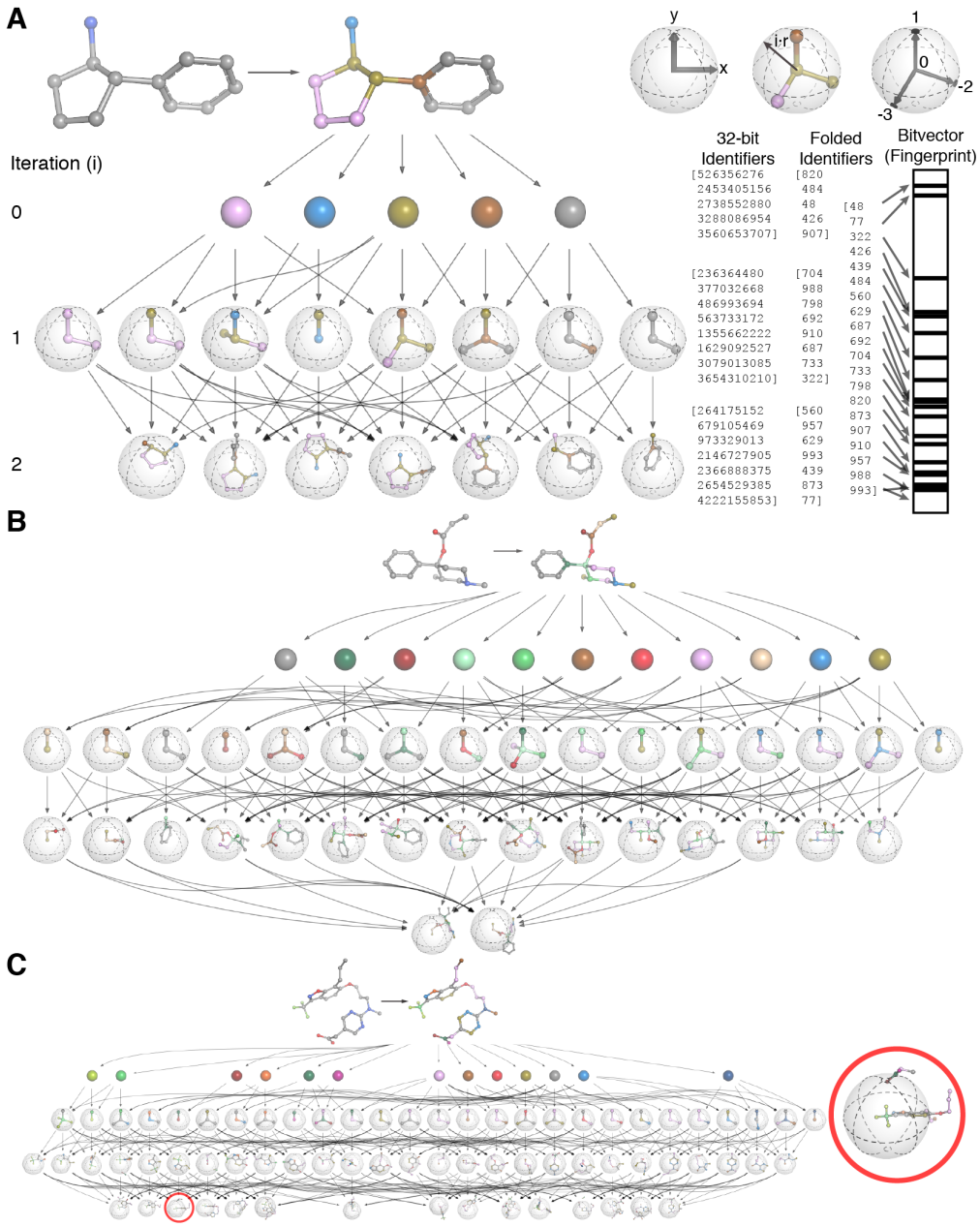

Figure 2.1: Diagram of information flow in the E3FP algorithm. A) Overview of fingerprinting process for cypenamine. At iteration 0, we assign atom identifiers using a list of atomic invariants and hash these into integers (shown here also as unique atom colors). At iteration i, shells of radius $i \cdot r$ center on each atom (top right). The shell contains bound and unbound neighbor atoms. Where possible, we uniquely align neighbor atoms to the xy-plane (top right) and assign stereochemical identifiers. Convergence occurs when a shell's substructure contains the entire molecule (third from the right) or at the maximum iteration count. Finally we "fold" each iteration's substructure identifiers to 1024-bit space. B) Overview of fingerprinting for compound **1**. C) Overview of fingerprinting for a large, flexible molecule (CHEMBL210990; expanded in Figure S1). A three-dimensional substructure can consist of two disconnected substructures and their relative orientations (right).

ochemistry). We designed this process to be minimally sensitive to minor structural fluctuations, so that conformers could be distinguished while the set of conformers for a given molecule would retain a degree of internal similarity in E3FP space. https://www.overleaf.com/project/606e379906d9bb5d62653c22 The binding-relevant conformers of most small molecules are not known *a priori*. Accordingly, prior to constructing any 3D fingerprint, we generate a library of potential conformers for the molecule, each of which in turn will have a unique fingerprint. We employed a previously published protocol using the open-source RDKit package[35], wherein the authors determined the number of conformers needed to recover the correct ligand conformation from a crystal structure as a function of the number of rotatable bonds in the molecule, with some tuning (see Experimental Section).

E3FP encodes small molecule 3D substructures

The core intuition of E3FP generation (**Fig. 2.1**a) is to draw concentrically larger shells and encode the 3D atom neighborhood patterns within each of them. To do so, the algorithm proceeds from small to larger shells iteratively. First, as in ECFP, we uniquely represent each type of atom and the most important properties of its immediate environment. To do so, we assign 32-bit integer identifiers to each atom unique to its count of heavy atom immediate neighbors, its valence minus neighboring hydrogens, its atomic number, its atomic mass, its atomic charge, its number of bound hydrogens, and whether it is in a ring. This can result in many fine-grained identifiers, some examples of which are visualized as differently colored atoms for the molecule cypenamine in **Fig. 2.1**a and for larger molecules in **Fig. 2.1**b-c.

At each subsequent iteration, we draw a shell of increasing radius around each atom, defining the neighbors as the atoms within the shell as described above. The orientation and connectivity of the neighbors–or lack thereof (as in **Fig. 2.1**c, red circle, expanded in Figure S1)–is combined with the neighbors' own identifiers from the previous iteration to generate a new joint identifier. Thus, at any given iteration, the information contained

within the shell is the union of the substructures around the neighbors from the previous iterations merged with the neighbors' orientation and connectivity with respect to the center atom of the current shell. The set of atoms represented by an identifier therefore comprise a three-dimensional substructure of the molecule.

We continue this process up to a predefined maximum number of iterations or until we have encountered all substructures possible within that molecule. We then represent each identifier as an "on" bit in a sparse bit vector representation of the entire conformer (**Fig. 2.1**a, bitvector). Each "on" bit indicates the presence of a specific three-dimensional substructure. The choice of numerical integer to represent any identifier is the result of a hash function (see Experimental Section) that spreads the identifiers evenly over a large integer space. Because there are over four billion possible 32-bit integers and we observe far fewer than this number of molecular substructures (identifiers) in practice, each identifier is unlikely to collide with another and may be considered unique to a single atom or substructure. Since this still remains a mostly empty identifier space, we follow the commonly used approach from ECFP, and "fold" E3FP down to a shorter bitvector for efficient storage and swift comparison; adapting the 1024-bit length that has been effective for ECFP4[6,36] (Table S2).

To demonstrate the fingerprinting process, **Fig. 2.1**a steps through the generation of an E3FP for the small molecule cypenamine. First, four carbon atom types and one nitrogen atom type are identified, represented by five colors. As cypenamine is fairly small, E3FP fingerprinting terminates after two iterations, at which point one of the substructures consists of the entire molecule. The slightly larger molecule **1** (CHEMBL270807) takes an additional iteration to reach termination (**Fig. 2.1**b). **Fig. 2.1**c and Figure S1 demonstrate the same process for CHEMBL210990. This molecule is more complex, with 13 distinct atom types, and in the conformation shown reaches convergence in three iterations. Because the molecule bends back on itself, in the second and third iterations, several of the identifiers represent substructures that are nearby each other in physical space but are not directly bound to each and other and indeed are separated by many bonds (*e.g.*, red circle in **Fig. 2.1**c). 2D finger-

prints such as ECFP are inherently unaware of unconnected proximity-based substructures, but they are encoded in E3FP.

SEA 3D fingerprint performance exceeds that of 2D in binding prediction

We were curious to determine how molecular similarity calculations using the new E3FP representations would compare to those using the 2D but otherwise similarly-motivated ECFP4 fingerprints. Specifically, we investigated whether the 3D fingerprint encoded information that would enhance performance over its 2D counterpart in common chemical informatics tasks.

The ECFP approach uses several parameters, (*e.g.*, ECFP4 uses a radius of 2), and prior studies have explored their optimization[36]. We likewise sought appropriate parameter choices for E3FP. In addition to the conformer generation choices described above, E3FP itself has four tunable parameters: 1) a shell radius multiplier (r in **Fig. 2.1**a), 2) number of iterations (i in Figure 1a), 3) inclusion of stereochemical information, and 4) final bitvector length (1024 in **Fig. 2.1**a). We explored which combinations of conformer generation and E3FP parameters produced the most effective 3D fingerprints for the task of recovering correct ligand binders for over 2,000 protein targets using the Similarity Ensemble Approach (SEA). SEA compares sets of fingerprints against each other using Tanimoto coefficients (TC) and determines a *p-value* for the similarity among the two sets; it has been used to predict drug off-targets[4,5,37,38], small molecule mechanisms of action[39–41], and adverse drug reactions[4,42,43]. For the training library, we assembled a dataset of small molecule ligands that bind to at least one of the targets from the ChEMBL database with an IC_{50} of 10 µM or better. We then generated and fingerprinted the conformers using each E3FP parameter choice, resulting in a set of conformer fingerprints for each molecule and for each target. We performed a stratified 5-fold cross-validation on a target-by-target basis by setting aside

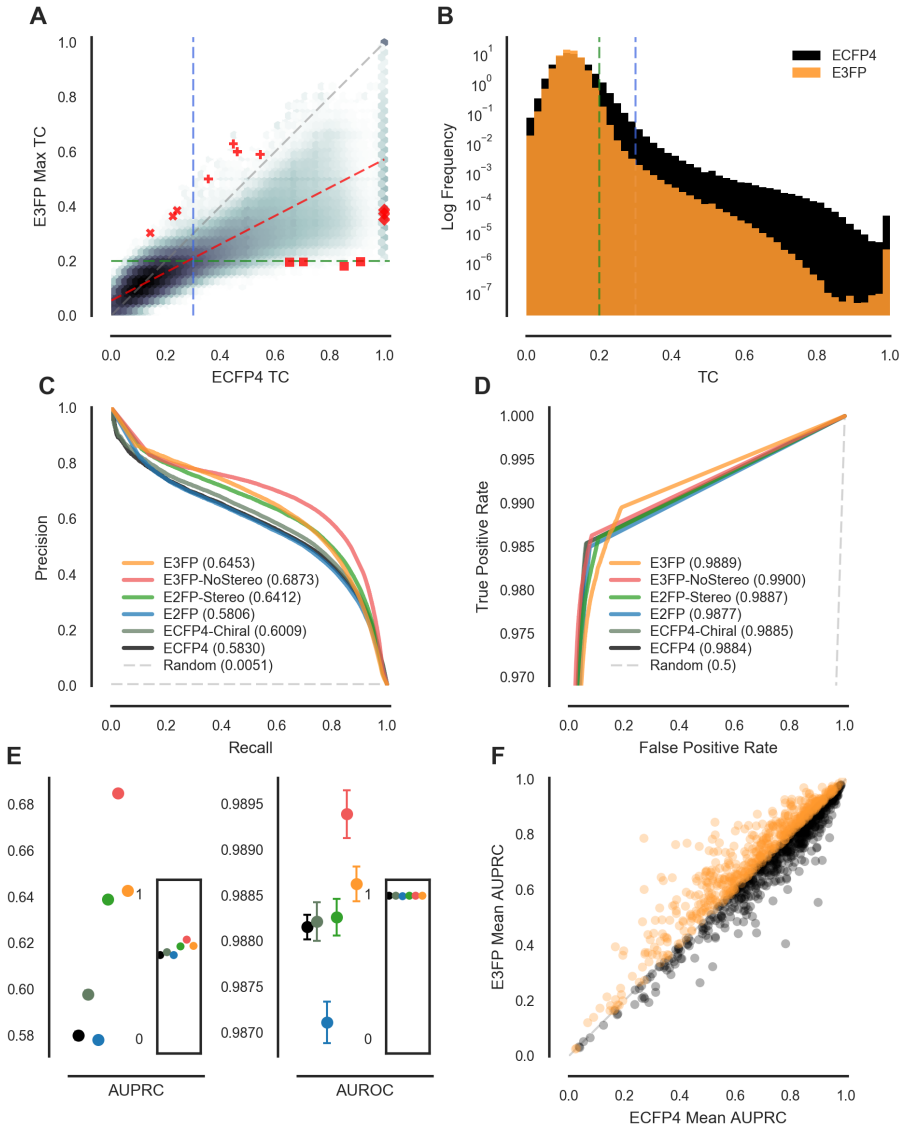

Figure 2.2: Comparative performance of E3FP and ECFP. For all pairs of 308,315 molecules from ChEMBL20, A) log density plot summarizing 95 billion maximum Tanimoto Coefficients (TC) calculated between E3FP conformer fingerprint sets versus corresponding TC by ECFP4 fingerprints. The dotted red line is a linear least squares fit. Optimal SEA TC cutoffs for E3FP (green) and ECFP4 (blue) are dotted lines. Red markers indicate examples in **Fig. 2.3**. B) Histograms of TCs from (A). C) Combined precision-recall (PRC) curves from 5 independent 5-fold cross- validation runs using 1024-bit E3FP, E3FP without stereochemical identifiers (E3FP-NoStereo), E3FP without stereochemical identifiers or nearby unbound atoms (E2FP), E3FP without nearby unbound atoms (E2FP-Stereo), ECFP4, and ECFP4 with distinct bond types encoding chirality (ECFP4-Chiral). Only the PRC of the highest AUC fold is shown. D) Combined highest-AUC ROC curves for the same sets as in (C). E) Results of bootstrapping AUCs as in **Table 2.1**. Dots indicate mean AUC, and whiskers standard deviations. Insets show absolute scale. F) Target-wise comparison of mean AUPRCs using E3FP versus ECFP4.

Table 2.1: Mean and standard deviations for combined fold AUPRC and AUROC curves versus target-wise AUPRC and AUROC curves across 5 independent repeats of 5-fold cross-validation are shown. A random classifier will produce a mean AUPRC of 0.0051 (fraction of positive target/mol pairs in test data), a mean target AUPRC of 0.0053 ± 0.0076, and a mean AUROC and mean target-wise AUROC of 0.5.

Name	Mean Fold AUPRC	Mean Fold AUROC	Mean Target AUPRC	Mean Target AUROC
ECFP4	0.5799 ± 0.0018	0.9882 ± 0.0001	0.6965 ± 0.2099	0.9772 ± 0.0387
ECFP4-Chiral	0.5977 ± 0.0017	0.9882 ± 0.0002	0.7021 ± 0.2088	0.9769 ± 0.0391
E2FP	0.5781 ± 0.0015	0.9871 ± 0.0002	0.7080 ± 0.2034	0.9768 ± 0.0392
E2FP-Stereo	0.6390 ± 0.0011	0.9883 ± 0.0001	0.7140 ± 0.2016	0.9780 ± 0.0371
E3FP-NoStereo	0.6849 ± 0.0012	0.9894 ± 0.0003	0.7312 ± 0.1989	0.9774 ± 0.0409
E3FP	0.6426 ± 0.0016	0.9886 ± 0.0002	0.7046 ± 0.1991	0.9805 ± 0.0326

one fifth of the known binders from a target for testing, searching this one fifth (positive data) and the remaining non-binders (negative data) against the target using SEA, and then computing true and false positive rates at all possible SEA *p-value* cutoffs. For each target in each fold, we computed the precision recall curve (PRC), the receiver operating characteristic (ROC), and the area under each curve (AUC). Likewise, we combined the predictions across all targets in a cross-validation fold to generate fold PRC and ROC curves.

As there are far more negative target-molecule pairs in the test sets than positives, a good ROC curve was readily achieved, as many false positives must be generated to produce a high false positive rate. Conversely, in such a case, the precision would be very low. We therefore expected the AUC of the PRC (AUPRC) to be a better assessment of parameter set[44]. To simultaneously optimize for both a high AUPRC and a high AUC of the ROC (AUROC), we used the sum of these two values as the objective function, $\mathrm{AUC}_{\mathrm{SUM}}$. We employed the Bayesian optimization program Spearmint[45] to optimize four of five possible E3FP parameters (we did not optimize fingerprint bit length, for simplicity of comparison to ECFP fingerprints) so as to maximize the $\mathrm{AUC}_{\mathrm{SUM}}$ value and minimize runtime of fingerprinting (Figure S2).

We constrained all optimization solely to the choice of fingerprint parameters, on the same

underlying collection of precomputed molecular conformers. For computational efficiency, we split the optimization protocol into two stages (see Experimental Section). This yielded an E3FP parameter set that used the three lowest energy conformers, a shell radius multiplier of 1.718, and 5 iterations of fingerprinting (Figure S4). After bootstrapping with 5 independent repeats of 5-fold cross-validation using E3FP, and ECFP4 on a larger set of 308,316 ligands from ChEMBL20, E3FP produced a mean AUPRC of 0.6426, exceeding ECFP4's mean AUPRC of 0.5799 in the same task (**Fig. 2.2**c,e; **Table 2.1**). Additionally, E3FP's mean AUROC of 0.9886 exceeds ECFP4's AUPRC of 0.9882 (**Fig. 2.2**d-e; **Table 2.1**). Thus, at a SEA *p-value* threshold $p \leq 3.45 \times 10^{-47}$, E3FP achieves an average *sensitivity* of 0.6976, *specificity* of 0.9974, *precision* of 0.5824, and F_1 score of 0.6348. ECFP4 achieves 0.4647, 0.9986, 0.6236, and 0.5325, at this *p-value* threshold. ECFP4 is unable to achieve the high F_1 score of E3FP, but at its maximum F_1 score of 0.5896 it achieves a *sensitivity* of 0.6930, a specificity of 0.9966, and a *precision* of 0.5131 using a *p-value* threshold $p \leq 3.33 \times 10^{-23}$. To ensure a fair comparison, we subjected ECFP to a grid search on its radius parameter and found that no radius value outperforms ECFP4 with both AUPRC and AUROC (Table S1). Additionally, fingerprints with longer bit lengths did not yield significant performance increases for E3FP or ECFP4, despite the expectation that longer lengths would lower feature collision rates (Table S2); indeed, it appears that increasing the fingerprint length reduced the performance of E3FP. By design, this optimization and consequent performance analysis does not attempt to quantify novelty of the predictions, nor assess the false negative or untested-yet-true- positive rate of either method.

We note that E3FP was optimized here for use with SEA, and SEA inherently operates on sets of fingerprints, such as those produced when fingerprinting a set of conformers. Most machine learning methods, however, operate on individual fingerprints. To determine how well E3FP could be integrated into this scenario, we repeated the entire cross-validation with four common machine learning classifiers: Naive Bayes Classifiers (NB), Random Forests (RF), Support Vector Machines with a linear kernel (LinSVM), and Artificial Neural Net-

works (NN). As these methods process each conformer independently, we computed the maximum score across all conformer-specific fingerprints for a given molecule, and used that score for cross-validation. Compared to cross-validation with SEA, LinSVM and RF produced better performance by PRC using both E3FP and ECFP4, while NB and RF suffered a performance loss (Figure S5). For ECFP4, this trend continued when comparing ROC curves, while for E3FP it did not (Figure S6). In general, the machine learning methods underperformed when using E3FP compared to ECFP4. When we instead took the bitwise mean of all conformer-specific E3FPs to produce one single summarizing "float" fingerprint per molecule, we observed an improvement across all machine learning methods except for LinSVM. The most striking difference was for RF, where performance with "mean E3FP" then matched ECFP4.

3D fingerprints encode different information than their 2D counterparts

2D fingerprints such as ECFP4 may denote stereoatoms using special disambiguation flags or identifiers from marked stereochemistry (here termed "ECFP4-Chiral")[32]. E3FP encodes stereochemistry more natively. Conceptually, all atoms within a spatial "neighborhood" and their relative orientations within that region of space are explicitly considered when constructing the fingerprint. To quantify how stereochemical information contributes to E3FP's improved AUPRC over that of ECFP4, we constructed three "2D- like" limited variants of E3FP, each of which omits some 3D information and is thus more analogous to ECFP. The first variant, which we term "E2FP," is a direct analogue of ECFP, in which only information from directly bound atoms are included in the identifier and stereochemistry is ignored. This variant produces similar ROC and PRC curves to that of ECFP4 (**Fig. 2.2**c-d; Figures S7-S8). A second variant, "E2FP-Stereo," includes information regarding the relative orientations of bound atoms. E2FP-Stereo achieves a performance between that of ECFP4

and E3FP, demonstrating that E3FP's approach for encoding stereochemical information is effective (**Fig. 2.2**c-d). The third variant, "E3FP-NoStereo," includes only the information from bound and unbound atoms. E3FP-NoStereo performs slightly better than E3FP in both ROC and PRC analysis (**Fig. 2.2**c-d), indicating that E3FP's enhanced performance over ECFP4 in PRC analysis is due not only to the relative orientations of atoms but also due to the inclusion of unbound atoms. All variants of E3FP with some form of 3D information outperformed both ECFP4 and ECFP4-Chiral (**Fig. 2.2**c-d; Figures S7-S8).

On average, the final E3FP parameters yield fingerprints with 35% more "on" bits than ECFP4, although if run for the same number of iterations, ECFP is denser. Thus E3FP typically runs for more iterations (Figure S4c-d). Folding E3FP down to 1024 bits results in an average loss of only 1.4 bits to collisions. The TCs for randomly chosen pairs of molecules by E3FP are generally lower (**Fig. 2.2**a-b) than those for ECFP4, and there are fewer molecules with identical fingerprints by E3FP than by ECFP4. The final E3FP parameter set outperforms ECFP up to the same number of iterations (Table S1, Figure 2c-d). Intriguingly, E3FP outperforms ECFP4 at this task on a per- target basis for a majority of targets (**Fig. 2.2**f).

Fourteen molecular pairs where 3D and 2D fingerprints disagree

To explore cases where E3FP and ECFP4 diverge, we computed E3FP versus ECFP4 pairwise similarity scores (Tanimoto coefficients; TCs) for all molecule pairs in ChEMBL20 (red markers in **Fig. 2.2**a). We then manually inspected pairs from four regions of interest. Pairs representative of overall trends were selected, with preference toward pairs that had been assayed against the same target (Table S3). The first region contains molecule pairs with TCs slightly above the SEA significance threshold for E3FP but below the threshold for ECFP4 (denoted by 'x' markers). These predominantly comprise small compact molecules, with common atom types across multiple orders or substituents on rings (**Fig. 2.3**a).

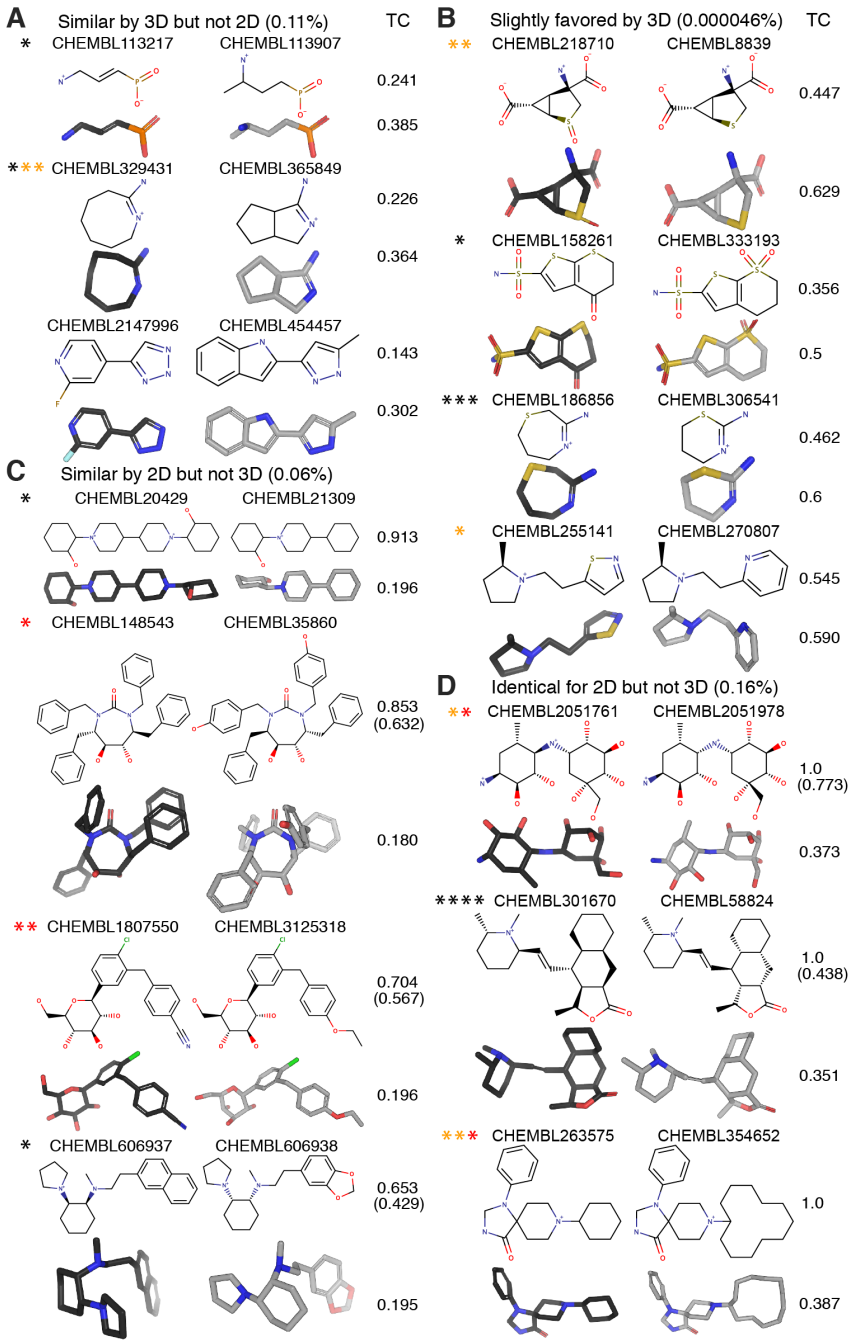

Figure 2.3: Examples of molecule pairs with high differences between E3FP and ECFP Tanimoto coefficients. Molecule pairs were manually selected from regions of interest, displayed as red markers in **Fig. 2.2**a: A) Upper left, B) Upper right, C) Lower right, and D) Far right. Pair TCs for ECFP4 and E3FP are shown next to the corresponding 2D and 3D representations; the conformer pairs shown are those corresponding to the highest pairwise E3FP TC. Where pair TCs for ECFP4 with stereochemical information differ from standard ECFP4, they are included in parentheses. Each colored asterisk indicates a target for which existing affinity data for both molecules was found in the literature and is colored according to fold-difference in affinity: black for <10-fold, orange for 10-100-fold, red for >100-fold.

Some of these molecules are already reported to interact with the same protein targets. For instance, CHEMBL113217 binds to GABA-B receptor with an IC_{50} of 280 nM, while CHEMBL113907 binds GABA-B with a similar IC_{50} of 500 nM (**Fig. 2.3**a)[46]. In another example, CHEMBL329431 binds to inducible, brain, and endothelial human nitric-oxide synthases with IC_{50}s of 10.0 µM, 10.1 µM, and 59 µM, respectively[47], while CHEMBL365849 binds to the same targets at 3.1 µM, 310 nM, and 4.7 µM[48]. The black asterisk alongside this pair marks similar affinities for the first target (within 1 log), and the gold asterisks affinities for the second two, each spanning two logs. Red asterisks mark targets whose affinities differ by more than two logs, but no such cases were found for this region.

The second region (red crosses in **Fig. 2.2**a) contains molecule pairs with TCs considered significant both in 2D and in 3D, but whose similarity was nonetheless greater by 3D (**Fig. 2.3**b). For instance, the molecule pairs often differed by atom types in or substituents on a ring, despite a high degree of similarity in 3D structures. In the case of CHEMBL158261 and CHEMBL333193, the molecules bind to carbonic anhydrase II with near-identical affinities of 3.6 nM and 3.3 nM[49]. Interestingly, the 2D similarity of this pair is barely above the significance threshold. In another example, the molecules CHEMBL186856 and CHEMBL306541 achieve markedly similar pharmacological profiles, as the first binds to the inducible, brain, and endothelial human nitric-oxide synthases with IC_{50}s of 1.2 µM, 2.8 µM, and 10.5 µM[50], whereas the second was reported at 2.9 µM, 3.2 µM, and 7.1 µM[51]. On the other hand, two other pairs somewhat differ in binding profile: while CHEMBL218710 binds to metabotropic glutamate receptors 2 and 3 with K_is of 508 nM and 447 nM, CHEMBL8839 binds to these targets more potently, at 40.6 nM and 4.7 nM[52]. Likewise, the binding profiles of CHEMBL255141 and **1** to histamine H3 receptor differed by approximately an order of magnitude, with respective K_is of 17 nM and 200 nM[53].

The third region (red squares in **Fig. 2.2**a) contains molecule pairs significant in 2D but not in 3D (**Fig. 2.3**c), and the fourth region (red diamonds in **Fig. 2.2**a) contains pairs identical by 2D yet dissimilar in 3D (**Fig. 2.3**d). These examples span several categories:

First, the conformer generation protocol failed for some pairs of identical or near-identical molecules having many rotatable bonds, because we generated an insufficient number of conformers to sample the conformer pair that would attain high 3D similarity between them (not shown). Second, in cases where the 2D molecules do not specify chirality, the specific force field used may favor different chiralities, producing artificially low 3D similarity. As an example, CHEMBL20429 and CHEMBL21309 (**Fig. 2.3**c) have relatively similar affinities for vesicular acetylcholine transporter at 200 nM and 40 nM[54] despite their low 3D similarity. Third, some pairs consist of molecules primarily differentiated by the size of one or more substituent rings (**Fig. 2.3**c-d). ECFP4 is incapable of differentiating rings with 5 or more identical atom types and only one substituent, while E3FP substructures may include larger portions of the rings. The role of ring size is revealed in the target affinity differences for one such pair: CHEMBL263575 binds to the kappa opioid, mu opioid, and nociceptin receptors with K_is of 100 nM, 158 nM, and 25 nM, while CHEMBL354652 binds to the same receptors notably more potently at 2.9 nM, 0.28 nM, and 0.95 nM[55]. Fourth, many pairs consist of molecules primarily differentiated by the order of substituents around one or more chiral centers (**Fig. 2.3**c-d). The molecules CHEMBL148543 and CHEMBL35860, for example, bind to HIV type 1 protease with disparate K_is of 560 nM[56] and 0.12 nM[57] despite their exceptionally high 2D similarity of 0.853 TC. Likewise, CHEMBL1807550 and CHEMBL3125318 have opposing specificities for the human sodium/glucose cotransporters 1 and 2; while the former has IC_{50}s of 10 nM and 10 µM for the targets[58], the latter has IC_{50}s of 3.1 µM and 2.9 nM[59]. In another example, despite being identical by standard 2D fingerprints, the stereoisomers CHEMBL2051761 and CHEMBL2051978 bind to maltase-glucoamylase with IC_{50}s of 28 nM versus 1.5 µM, and to sucrase-isomaltase at 7.5 nM versus 5.3 µM[60]. The stereoisomers CHEMBL301670 and CHEMBL58824, however, show a case where 3D dissimilarity is a less effective guide, as both molecules bind to the muscarinic acetylcholine receptors M_1-M_4 with generally similar respective IC_{50}s of 426.58 nM versus 851.14 nM, 95.5 nM versus 851.14 nM, 1.6 µM versus 794.33 nM, and 173.78 nM versus 794.33 nM[61]. Similarly,

CHEMBL606937 and CHEMBL606938 have low similarity in 3D but bind to sigma opioid receptor with IC_{50}s of 37 and 34 nM[62].

E3FP predicts correct new drug off-targets that are not apparent in 2D

As E3FP enhanced SEA performance in retrospective tests (**Fig. 2.2**c-d), we hypothesized that this combination might identify novel interactions as yet overlooked with two-dimensional fingerprints. We therefore tested whether SEA with E3FP would make correct drug-to-target predictions that SEA with ECFP4 did not make. Using a preliminary choice of E3FP parameters (Table S4), we generated fingerprints for all in-stock compounds in the ChEMBL20 subset of the ZINC15 (`zinc15.docking.org`) database with a molecular weight under 800 Da. As our reference library, we extracted a subset of ChEMBL20 comprising 309 targets readily available for testing by radioligand binding assay in the Psychoactive Drug Screening Program (PDSP)[63] database. Using SEA on this library, we identified all drug-to-target predictions with a *p-value* stronger than 1×10^{-25}. To focus on predictions specific to E3FP, we removed all predictions with a *p-value* stronger than 0.1 when counter-screened by SEA with ECFP4, resulting in 9,331 novel predicted interactions. We selected eight predictions for testing by binding assay; of these, five were inconclusive, and three bound to the predicted target subtype or to a close subtype of the same receptor (Table S4-S7). We address each of the latter in turn.

The E3FP SEA prediction that the psychostimulant and antidepressant[64–66] cypenamine (CHEMBL2110918, KEGG:D03629), for which we could find no accepted targets in the literature despite its development in the 1940s, would bind to the human nicotinic acetylcholine receptor (nAchR) $\alpha 2\beta 4$ was borne out with a K_i of 4.65 µM (**Fig. 2.4**c; Table S7). Of note, this corresponds to a high ligand efficiency (LE) of 0.610 kcal/mol/heavy atom (see Experimental Section). An LE greater than 0.3 kcal/mol/heavy atom atom is generally considered

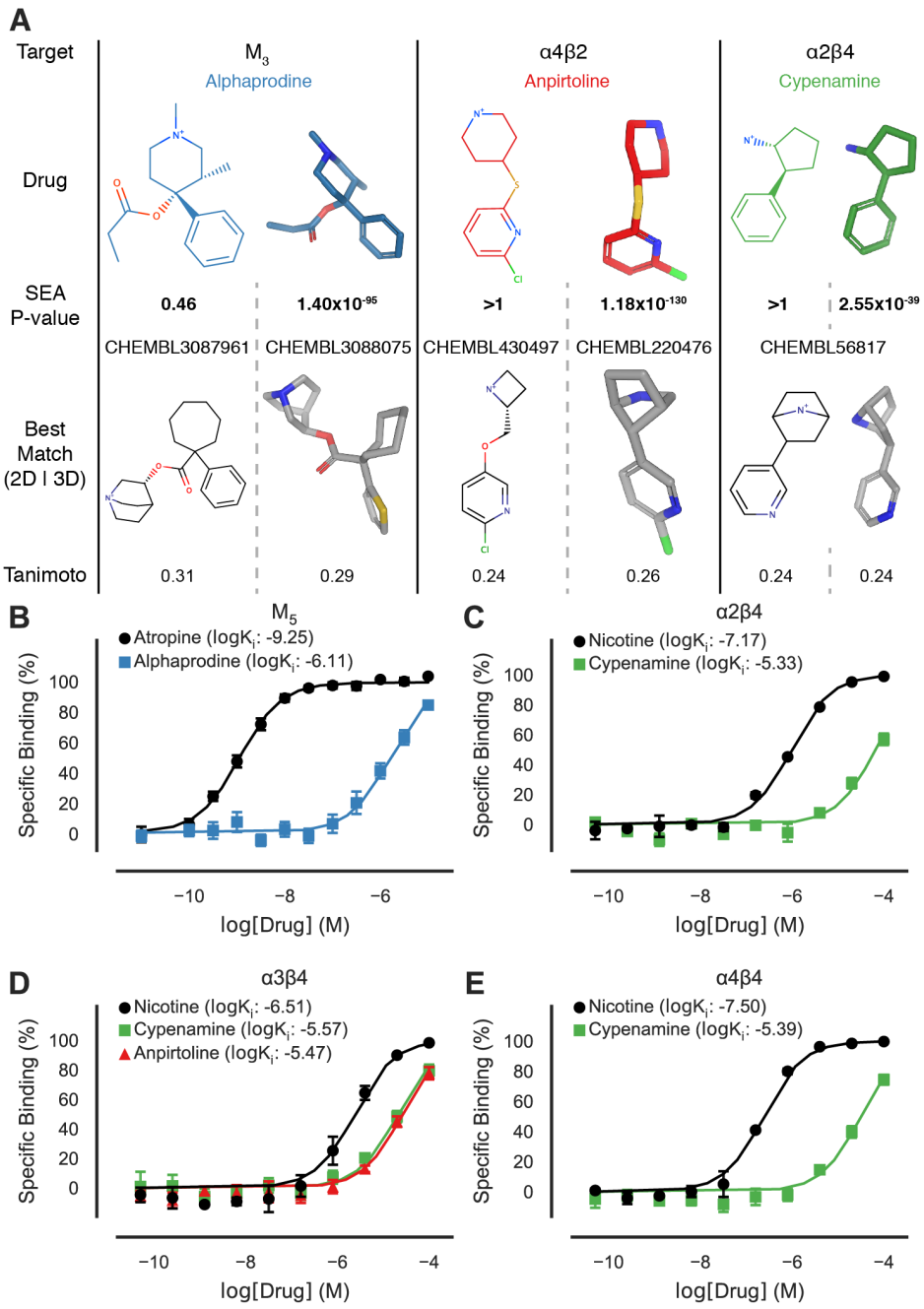

Figure 2.4: Experimental results of novel compound-target predictions. A) SEA predictions that motivated the binding experiments, with 2D versus 3D SEA *p-values* for each drug-target pair. Tanimoto coefficients score the similarity of 2D versus 3D structures for the searched drug against its most similar known ligand(s) of the target by ECFP (left) and E3FP (right). E3FP uses an early parameter set. Supporting Table S4 shows recalculated SEA *p-values* on the final E3FP parameter set used elsewhere. B-E) Experimentally measured binding curves for tested drugs and reference binders (black) at protein targets B) M_5, C) $\alpha 2\beta 4$, D) $\alpha 3\beta 4$, and E) $\alpha 4\beta 4$. See Table S7 for more details.

a promising drug candidate[67]. As any prediction is only as specific as the reference ligand data from ChEMBL upon which it was based, we assayed cypenamine against multiple subtypes of nAchR. Cypenamine also bound to the nAchR subtypes $\alpha3\beta4$ and $\alpha4\beta4$ with K_i's of 2.69 and 4.11 µM (**Fig. 2.4**d-e, Table S7) and ligand efficiencies of 0.637 and 0.616 kcal/mol/heavy atom.

Anpirtoline (CHEMBL1316374) is an agonist of the 5-HT_{1B}, 5-HT_{1A}, and 5-HT_2 receptors, and an antagonist of the 5-HT_3 receptor, with K_i's of 28, 150, 1490, and 30 nM, respectively[68,69]. However, we predicted it would bind to the nAchRs, of which it selectively bound to $\alpha3\beta4$ at a K_i of 3.41 µM and an LE of 0.536 kcal/mol/heavy atom (**Fig. 2.4**d, Table S7). In this case, the motivating SEA E3FP prediction was for the $\alpha4\beta2$ subtype of nAchR, for which the experiment was inconclusive, suggesting either that the ligand reference data from ChEMBL distinguishing these subtypes was insufficient, or that the SEA E3FP method itself did not distinguish among them, and this is a point for further study.

Alphaprodine (CHEMBL1529817), an opioid analgesic used as a local anesthetic in pediatric dentistry[70], bound to the muscarinic acetylcholine receptor (mAchR) M_5 with a K_i of 771 nM and an LE of 0.442 kcal/mol/heavy atom (**Fig. 2.4**b, Figure S9e). We found no agonist activity on M_5 by alphaprodine by Tango assay[71,72] (Figure S10b), but we did find it to be an antagonist (Figure S11). Intriguingly, alphaprodine also showed no significant affinity for any of the muscarinic receptors M_1-M_4 up to 10 µM (Figures S9a-d), indicating that it is an M_5-selective antagonist. Muscarinic M_5 selective small molecules are rare in the literature[73]. Whereas its M_5 selectivity would need to be considered in the context of its opioid activity (mu, kappa, and delta opioid receptor affinities however are not publicly available), alphaprodine nonetheless may find utility as a M_5 chemical probe, given the paucity of subtype-selective muscarinic compounds. Interestingly, the E3FP SEA prediction leading us to the discovery of this activity was for the muscarinic M_3 receptor, to which alphaprodine ultimately did not bind and for which alphaprodine showed no agonist activity (Figure S10a). This highlights not only the limitations of similarity-based methods such as SEA for

the discovery of new subtype-selective compounds when none of that type are previously known, but also the opportunity such methods provide to identify chemotypes and overall receptor families that merit further study nonetheless.

2.4 Discussion

Three results emerge from this study. First, we encode a simple three-dimensional molecular representation into a new type of chemical informatic fingerprint, which may be used to compare molecules in a manner analogous to that already used for two-dimensional molecular similarity. Second, the 3D fingerprints contain discriminating information that is naturally absent from 2D fingerprints, such as stereochemistry and relationships among atoms that are close in space but distant in their direct bond connectivity. Finally, as small molecules may adopt many structural conformations, we combine conformation-specific 3D fingerprints into sets to evaluate entire conformational ensembles at once. This may be of interest in cases where different conformations of a molecule are competent at diverse binding sites across the array of proteins for which that same molecule is, at various potencies, a ligand.

We devised a simple representation of three-dimensional molecular structures, an "extended 3D fingerprint" (E3FP), that is directly analogous to gold standard two-dimensional approaches such as the extended connectivity fingerprint (ECFP). As with two-dimensional fingerprints, this approach enables pre-calculation of fingerprints for all conformers of interest in an entire library of molecules once. Unlike conventional 3D approaches, similarity calculations in E3FP do not require an alignment step. Consequently, E3FP similarity calculations are substantially faster than standard 3D comparison approaches such as ROCS. Furthermore, E3FP fingerprints are formatted identically to ECFP and other 2D fingerprints. Thus systems pharmacology approaches such as SEA[4,5], Naive Bayes Classifiers[9], SVM[14], and other established machine learning methods may readily incorporate E3FPs for molecular conformers without modification. While choices of E3FP's parameter space might be specif-

ically optimized for the machine learning method in question, we have demonstrated that E3FP's highest-performing parameter choice for SEA (**Fig. 2.2**c-d) produces fingerprints that likewise perform well for SVM, random forests, and neural networks (Figures S5-S6).

To explore the role of 2D vs 3D features in the discriminatory power of molecular fingerprints, we progressively disabled capabilities specific to E3FP, such as stereochemistry encoding (termed "E3FP-NoStereo") and non-bonded atom relationships (termed "E2FP-Stereo"), eventually arriving at a stripped-down version of E3FP (termed "E2FP") that, much like ECFP, encodes only 2D information. We evaluated the consequences of removing these three-dimensional features on performance in retrospective machine learning tasks (*e.g.*, **Fig. 2.2**c-e; **Table 2.1**; Figures S7-S8) We found that inclusion of non-bonded atoms was a more important contributor to performance than stereochemical information. Intriguingly, while progressively adding stereochemical information and inclusion of nonbonded atoms produces marked improvement over ECFP4, inclusion only of nonbonded atom information produces the highest performance fingerprint of all, perhaps because 3D orientations of larger substructures are implicitly encoded within shells purely by relative distances. This observation leads us to believe that a more balanced inclusion of stereochemical information and nonbonded atoms may produce an even higher performing fingerprint. Historically, 3D representations have typically underperformed 2D ones such as ECFP[7], and this has always been the case with Similarity Ensemble Approach (SEA) calculations in particular[6]. Here, however, we find that E3FP exceeds the performance of ECFP in its precision-recall curve (PRC) and matches that of ECFP in its receiver-operating characteristic curve (ROC) area under the curve (AUC) scores (**Fig. 2.2**c-e; **Table 2.1**; Figures S7-S8). While the ROC curve evaluates the general usefulness of the fingerprint for classification by comparing sensitivity and specificity, the precision-recall evaluates how useful the method is for real cases where most tested drug-target pairs are expected to have no affinity. The increased performance in PRC curves when using E3FP over ECFP4 therefore indicates an increased likelihood of correctly predicting novel drug-target pairs with no loss in predictive power.

E3FP's utility for this task became especially clear when we used it to predict novel drug to protein binding interactions. To do so, we examined only strong SEA predictions with E3FP (SEA-E3FP; *p-value* $\leq 1 \times 10^{-25}$) that could not be predicted using SEA with ECFP (SEA-ECFP; *p-value* ≤ 0.1). We considered this a challenging task because on-market drugs might be expected to have fewer unreported off-targets in general than a comparatively newer and less-studied research compound might. Furthermore, much of the prior work in chemical informatics guiding molecule design and target testing has been motivated by 2D approaches[2,7,74]. Accordingly, approximately half of the new predictions were inconclusive in this first prospective test of the method (Tables S4 and S6). Nonetheless, many also succeeded with high ligand efficiencies (LEs), and these included unique selectivity profiles (**Fig. 2.4**). In one example, SEA-E3FP successfully predicted that alphaprodine would also act as an antagonist of the M_5 muscarinic receptor, which to our knowledge is not only a new "off-target" activity for this drug, but also constitutes a rare, subtype selective M_5 antimuscarinic ligand[73]. The M_5 muscarinic receptor has roles in cocaine addiction[75], morphine addiction[76], and dilation of cerebral blood vessels, with potential implications for Alzheimer's disease[77]. Study of M_5 receptors has been hindered by a lack of selective ligands. Due to serious adverse reactions[78], alphaprodine was withdrawn from the market in the United States in 1986 and is therefore unlikely to be applied as a therapeutic. However, alphaprodine might be explored not only as a chemical probe for studying M_5, but also as a reference for future therapeutic development.

Anpirtoline and cypenamine, likewise predicted and subsequently experimentally confirmed to bind previously unreported off-targets among the nicotinic receptors, exhibited exceptional LEs (0.536 - 0.637 kcal/mol/heavy atom), a commonly used metric of optimization potential. Recent patents combining psychostimulants with low-dose antiepileptic agents for the treatment of attention deficit hyperactivity disorder (ADHD) incorporate cypenamine[79,80], and nicotinic agents improve cognition and combat ADHD[81]. Given likewise the association of nicotinic acetylcholine receptor (nAchR) $\alpha 4$ gene polymorphisms

with ADHD[82], a combination of traditional psychostimulant activity with "non-stimulant" nAChR activity via $\alpha 4$ might improve anti-ADHD efficacy. Whereas cypenamine's micromolar binding concentration to nAchR is likely below the plasma concentrations it reaches at steady state, its exceptional LEs at nAchR may support further optimization of this pharmacology. As with cypenamine, anpirtoline may serve as a well-characterized starting point for further nAchR optimization, and secondarily, its serotonergic activity may serve as a guide to explore cypenamine's likely serotonergic activity. Anpirtoline's benign side effect profile, combined with the nAchR $\alpha 3\beta 4$ subunit's role in nicotine addiction[83] and the lack of $\alpha 3\beta 4$ specific drugs[84] , motivate further exploration.

We find that, whereas E3FP's performance matches or exceeds that of ECFP under multiple retrospective metrics, and whereas it leads to new off-target predictions complementing those of ECFP with SEA, there are cases where the more traditional 2D representation yields higher retrospective performance. It would be difficult to tease out the impact that 2D has of necessity made in guiding the design and testing of such molecules, and only time will tell whether ECFP's higher performance in these cases is due to true pharmacology or historical bias. However, we currently find that ECFP outperforms E3FP on specific targets using SEA (**Fig. 2.2**f) and in general when applying other machine learning methods (Figures S5-S6). Similarly, ECFP performs well on highly flexible molecules, owing to the difficulty of a small conformer library representing the flexibility of these molecules. Conversely, E3FP's potential for discerning similar target binding profiles is best realized when comparing molecules with a high degree of conformational similarity on the one hand or on the other one or more chiral centers. As is evident from their respective PRC plots, E3FP typically discriminates SEA predictions more than ECFP does, thereby achieving a better precision-recall ratio, at the initial cost of some sensitivity (**Fig. 2.2**c). However, this also allows E3FP to consider more distant molecular similarity relationships while maintaining greater discriminatory power than ECFP does at this range. It would be interesting to explore whether some of these more distant relationships might also be regions of pharmacological novelty.

One longtime advantage of 2D molecular representations has been their ability to implicitly sidestep the question of conformation. Whereas heroic effort has gone into solving the crystallographic conformations of hundreds of thousands of small molecules[85,86], the binding-competent 3D conformations for millions of research[25] and purchasable[31] small molecules are not known. Furthermore, polypharmacology exacerbates this problem, wherein a single small molecule can bind many protein partners, as it is not always the case that the molecule in question will adopt the same conformation for each binding site[2]. Powerful methods to enumerate and energetically score potential conformations exist[87–89], but it falls to the researcher to prioritize which of these conformers may be most relevant for a given protein or question. Treating the top five, ten, or more most energetically favorable conformers as a single set, however, may be an alternate solution to this problem. We originally developed SEA so as to compare entire sets of molecular fingerprints against each other[4], so it seemed natural to use it in a conformational-set-wise manner here. Furthermore, because SEA capitalizes on nearest-neighbor similarities among ligands across sets of molecules, we expected that it might analogously benefit from nearest-neighbor similarities in conformational space, on a protein-by-protein basis. This may indeed be the case, although we have not attempted to deconvolve E3FP's performance in a way that would answer whether different E3FPs, and hence different conformations, of the same molecule most account for its predicted binding to different protein targets.

The E3FP approach is not without its limitations. E3FP fingerprints operate on a pre-generated library of molecular conformers. The presence of multiple conformers and therefore multiple fingerprints for a single molecule hampers machine learning performance in naive implementations (Figures S5-S6), as flexible molecules dominate the training and testing data. We anticipate higher numbers of accepted conformers to only exacerbate the problem. The full conformational diversity of large, flexible molecules pose a substantial representational challenge as well (**Fig. 2.3**c-d). As E3FP depends upon conformer generation, a generator that consistently imposes specific stereochemistry on a center lacking chiral infor-

mation may produce artificially low or high 3D similarity (**Fig. 2.3**c). Furthermore, the core intuition of E3FP hinges on the assumption that most binding sites will have differing affinities for molecules with diverging stereochemical orientations, such as stereoisomers. Due to site flexibility, this is not always the case (**Fig. 2.3**c-d).

Despite these caveats, we hope that this simple, rapid, and conformer-specific extended three-dimensional fingerprint (E3FP) will be immediately useful to the broader community. To this end, we have designed E3FP to integrate directly into the most commonly used protein target prediction methods without modification. An open-source repository implementing these fingerprints and the code to generate the conformers used in this work is available at `https://github.com/keiserlab/e3fp/tree/1.0`.

2.5 Experimental Section

Generating Conformer Libraries

To maximize reproducibility, we generated conformers following a previously published protocol[35] using RDKit[90]. For each molecule, the number of rotatable bonds determined the target number of conformers, N, such that: $N = 50$ for molecules with less than 8 rotatable bonds, $N = 200$ for molecules with 8 to 12 rotatable bonds, and $N = 300$ for molecules with over 12 rotatable bonds. We generated a size $2N$ pool of potential conformers.

After minimizing conformers with the Universal Force Field[89] in RDKit, we sorted them by predicted energy. The lowest energy conformer became the seed for the set of accepted conformers. We considered each candidate conformer in sorted order, calculated its root mean square deviation (RMSD) to the closest accepted conformer, and added the candidate to the accepted set if its RMSD was beyond a predefined distance cutoff R. Optionally, we also enforced a maximum energy difference E between the lowest and highest energy accepted conformers. After having considered all $2N$ conformers, or having accepted N

conformers, the process terminated, yielding a final set of conformers for that molecule.

We tuned this protocol using three adjustable parameters: (1) the minimum root mean square distance (RMSD) between any two accepted conformers, (2) the maximum computed energy difference between the lowest energy and highest energy accepted conformers, and (3) the number of lowest energy conformers to be accepted (fingerprinted). We generated two different conformer libraries by this protocol. In the first (rms0.5), we used a RMSD cutoff $R = 0.5$, with no maximum energy difference E. In the second (rms1_e20), we chose a RMSD cutoff $R = 1.0$, with a maximum energy difference of 20 kcal/mol.

Enumerating Protonation States

Where specified, we generated dominant tautomers at pH 7.4 from input SMILES using the CXCALC program distributed with ChemAxon's Calculator Plugins[91]. We kept the first two protonation states with at least 20% predicted occupancy. Where no states garnered at least 20% of the molecules, or where protonation failed, we kept the input SMILES unchanged. Conformer generation for each tautomer proceeded independently and in parallel.

ECFP Fingerprinting

To approximate ECFP fingerprints, we employed the Morgan fingerprint from RDKit using default settings and an appropriate radius. ECFP4 fingerprints, for example, used a Morgan fingerprint of radius 2. Where ECFP with stereochemical information is specified, the same fingerprinting approach was used with chirality information incorporated into the fingerprint.

E3FP Fingerprinting

Given a specific conformer for a molecule, E3FP generates a 3D fingerprint, parameterized by a shell radius multiplier r and a maximum number of iterations (or level) L, analogous

to half of the diameter in ECFP. E3FP explicitly encodes stereochemistry.

Generating Initial Identifiers

Like ECFP, E3FP generation is an iterative process and can be terminated at any iteration or upon convergence. At iteration 0, E3FP generation begins by determining initial identifiers for each atom based on six atomic properties, identical to the invariants described in[32] : the number of heavy atom immediate neighbors, the valence minus the number of neighboring hydrogens, the atomic number, the atomic mass, the atomic charge, the number of bound hydrogens, and whether the atom is in a ring. For each atom, the array of these values are hashed into a 32-bit integer, the atom identifier at iteration 0. While the hashing function is a matter of choice, so long as it is uniform and random, this implementation used MurmurHash3[92].

Generating Atom Identifiers at Each Iteration

At each iteration i where $i > 0$, we consider each atom independently. Given a center atom, the set of all atoms within a spherical shell of radius $i \cdot r$ centered on the atom defines its immediate neighborhood, where the parameter r is the shell radius multiplier (**Fig. 2.1**a). We initialize an array of integer tuples with a number pair consisting of the iteration number i and the identifier of the central atom from the previous iteration.

For each non-central atom within the shell, we add to the array an integer 2-tuple consisting of a connectivity identifier and the atom's identifier from the previous iteration. The connectivity identifiers are enumerated as an expanded form of those used for ECFP: the bond order for bond orders of 1-3, 4 for aromatic bonds, and 0 for neighbors not bound to the central atom. To avoid dependence on the order in which atom tuples are added to the array, we sort the positions of all but the first tuple in ascending order. 3-tuples are then formed through the addition of a stereochemical identifier, followed by re-sorting. This

process is described in detail below.

We then flatten the completed array into a one-dimensional integer array. We hash this 1D array into a single new 32-bit identifier for the atom and add it to an identifier list for the iteration, after optional filtering described below.

Adding Stereochemical Identifiers

We generate stereochemical identifiers by defining unique axes from the sorted integer 2-tuples from the previous step combined with spatial information. First, we determine the vectors from the center atom to each atom within the shell. Then, we select the first unique atom by atom identifier from the previous iteration, if possible, and set the vector from the central atom to it as the y-axis. Where this is not possible, we set the y-axis to the average unit vector of all neighbors. Using the angles between each unit vector and the y-axis, the atom closest to 90 degrees from the y-axis with a unique atom identifier from the previous iteration defines the vector of the x-axis (**Fig. 2.1**a).

We then assign integer stereochemical identifiers s. Atoms in the $y > 0$ and $y < 0$ hemispheres have positive and negative identifiers, respectively. $s = \pm 1$ is assigned to atoms whose unit vectors fall within 5 degrees of the y-axis. We divide the remaining surface of the unit sphere into eight octants, four per hemisphere. The x-axis falls in the middle of the $s = \pm 2$ octants, and identifiers $\pm 3 - 5$ denote remaining octants radially around the y-axis (**Fig. 2.1**a). If unique y- and x-axes assignment fails, all stereochemical identifiers are set to 0.

Combining the connectivity indicator and atom identifier with the stereochemical identifier forms a 3-tuple for each atom, which, when hashed, produces an atom identifier dependent orientation of atoms within the shell.

Removing Duplicate Substructures

Each shell has a corresponding *substructure* defined as the set of atoms whose information is contained within the atoms in a shell. It includes all atoms within the shell on the current iteration as well as the atoms within their substructures in the previous iteration. Two shells have the same substructure when these atom sets are identical, even when the shell atoms are not. As duplicate substructures provide little new information, we filter them by only adding the identifiers to that iteration's list that correspond to new substructures or, if two new identifiers correspond to the same substructure, the lowest identifier.

Representing the Fingerprint

After E3FP runs for a specified number of iterations, the result is an array of 32-bit identifiers. We interpret these as the only "on" bits in a 2^{32} length sparse bitvector, and they correspond to 3D substructures. As with ECFP, we "fold" this bitvector to a much smaller length such as 1024 by successively splitting it in half and conducting bitwise OR operations on the halves. The sparseness of the bitvector results in a relatively low collision rate upon folding.

Fingerprint Set Comparison with SEA

The similarity ensemble approach (SEA) is a method for searching one set of bitvector fingerprints against another set[4]. SEA outputs the maximum Tanimoto coefficient (TC) between any two fingerprint sets and a *p-value* indicating overall similarity between the sets. SEA first computes all pairwise TCs between the two fingerprint sets. The sum of all TCs above a preset pairwise TC threshold T defines a *raw score*. For a given fingerprint, SEA calculates a background distribution of raw scores empirically[4]. This yields an observed *z-score* distribution, which at suitable values of T follows an extreme value distribution (EVD). For values of T ranging from 0 to 1, comparing goodness of fit (*chi-square*) to an

EVD vs a normal distribution determines an optimal range of T, where the empirical *z-score* distribution favors an EVD over a normal distribution. In this EVD regime we may convert a *z-score* to a *p-value* for any given set-set comparison.

K-fold Cross-Validation with SEA

We performed k-fold cross-validation on a target basis by dividing the ligands of at least µM affinity to each target into k sets per target. For a given fold, $k-1$ ligand sets and their target labels together formed the training data. The remaining ligand sets and their target labels formed the test data set. Due to the high number of negative examples in the test set, this set was reduced by ~25% by removing all negative target-molecule pairs that were not positive to any target in the test set. Conformers of the same ligand did not span the train vs test set divide for a target. For each fold, conformer fingerprint sets for molecules specific to the test set were searched against the union of all training conformer fingerprints for that target, yielding a molecule-to-target SEA *p-value*. From the -log *p-values* for all test-molecule-to-potential-target tuples, we constructed a receiving operator characteristic (ROC) curve for each target, and calculated its area under the curve (AUC). We likewise calculated the AUC for the Precision-Recall Curve (PRC) at each target. For a given fold, we constructed an ROC curve and a PRC curve using the -log *p-values* and true hit/false hit labels for all individual target test sets, which we then used to compute a fold AUROC and AUPRC. We then computed an average AUROC and AUPRC across all k folds. The objective function $\mathrm{AUC}_{\mathrm{SUM}}$ consisted of the sum of the average AUROC and AUPRC.

Optimizing Parameters with Spearmint

E3FP fingerprints have the following tunable parameters: stereochemical mode (on/off), nonbound atoms excluded, shell radius multiplier, iteration number, and folding level. Additional tunable parameters for the process of conformer generation itself are the minimum

RMSD between conformers, the maximum energy difference between conformers, and how many of the first conformers to use for searching. This parameter space forms a 8-dimensional hypercube. Of the 8 dimensions possible, we employed the Bayesian optimization program Spearmint[45] to explore four: shell radius multiplier, iteration number, number of first conformers, and two combinations of values for the RMSD cutoff and maximum energy difference between conformers. We evaluated the parameter sets by an objective function summing ROC and PRC AUCs (AUC_{SUM}), and Spearmint proposed future parameter combinations. The objective function evaluated k-fold cross-validation with the similarity ensemble approach (SEA) as described in the following section.

For the first stage, the dataset consisted of 10,000 ligands randomly chosen from ChEMBL17, the subset of targets that bound to at least 50 of these ligands at µM or better, and the objective function used was the AUPRC. Spearmint explored values of the shell radius multiplier between 0.1 and 4.0 Å, the number of lowest energy conformers ranging from 1 to all, and maximum iteration number of 5. Additionally, two independent conformer libraries were explored: rms0.5 and rms1_e20 (see above). 343 unique parameter sets were explored. We found that the best parameter sets used less than 35 of the lowest energy conformers, a shell radius multiplier between 1.3 and 2.8 Å, and 2-5 iterations. The conformer library used did not have an apparent effect on performance (data not shown).

For the second stage, we ran two independent Spearmint trajectories with a larger dataset consisting of 100,000 ligands randomly chosen from ChEMBL20, the subset of targets that bound to at least 50 of these ligands at µM or better, and the AUC_{SUM} objective function. We employed the CXCALC program[91] to determine the two dominant protonation states for each molecule at physiological pH, and then conformers were generated using an RMSD cutoff of 0.5. The number of fingerprinting iterations used in both trajectories was optimized from 2 to 5, but the two trajectories explored different subsets of the remaining optimal parameter ranges identified during the first stage: the first explored shell radius multipliers between 1.3 and 2.8 Å with number of conformers bounded at 35, while the second explored shell radius

multipliers between 1.7 and 2.8 Å with number of conformers bounded at 20. Spearmint tested 100 parameter combinations in each trajectory.

During optimization, we observed that the simple heuristic used by SEA to automatically select the TC threshold for significance resulted in folds with high TC cutoffs having very high AUPRCs but low AUROCs due to low recall, while folds with low TC cutoffs had lower AUPRCs but very high AUROCs (Figure S3). Several folds in the latter region outperformed ECFP4 in both AUPRC and AUROC (Figure S3c). We therefore selected the best parameter set as that which produced the highest AUC_{SUM} while simultaneously outperforming ECFP4 in both metrics. For all future comparisons, the TC cutoff that produced the best fold results was applied to all folds during cross-validation.

K-fold Cross-Validation with Other Classifiers

We performed k-fold cross-validation using alternative classifiers in the same manner as for SEA, with the following differences. We trained individual classifiers on a target by target basis. In the training and test data, we naively treated each conformer fingerprint as a distinct molecular fingerprint, such that the conformer fingerprints did not form a coherent set. After evaluating the target classifier on each fingerprint for a molecule, we set the molecule score to be the maximum score of all of its conformer fingerprints.

For the Naive Bayes (NB), Random Forest (RF), and Support Vector Machine with a linear kernel (LinSVM) classifiers, we used Scikit- learn version 0.18.1 (`https://github.com/scikit-learn/scikit-learn/tree/0.18.1`). We used default initialization parameters, except where otherwise specified. For the RF classifier, we used 100 trees with a maximum depth of 2. We weighted classes (positive and negative target/molecule pairs) to account for class imbalance. For LinSVM kernel, we applied an l1 norm penalty and balanced class weights as for RF.

We implemented Artificial Neural Network (NN) classifiers with nolearn version 0.6.0 (`https://github.com/dnouri/nolearn/tree/0.6.0`). We trained networks independently for each target using 1024-bit input representations from either E3FP or ECFP. The NN architecture comprised 3 layers: an input layer, a single hidden layer with 512 nodes, and an output layer. We used dropout[93] as a regularizer on the input and hidden layers at rates of 10% and 25%, respectively. The hidden layer activation function was Leaky Rectified Linear[94] with default leakiness of 0.01. The prediction layer used softmax nonlinearities. We trained networks trained for 1000 epochs with early stopping to avoid overfitting, by monitoring the previous 75 epochs for lack of change in the loss function. The final softmax layer contained 2 tasks (classes), one corresponding to binding and the other corresponding to the absence of binding. This softmax layer produced a vector corresponding to the probability of a given molecule being a binder or non-binder given the neural network model. We calculated training error using a categorical cross entropy loss.

Predicting Novel Compound-Target Binding Pairs

To identify novel compound-target pairs predicted by E3FP but not by ECFP, we built a subset of 309 proteins/complex mammalian targets (106 human) for which the National Institute of Mental Health Psychoactive Drug Screening Program (NIMH PDSP)[63] had established binding assays. We selected all compounds listed as in-stock in ZINC15[31], downloaded on 2015-09-24. We fingerprinted all ligands in ChEMBL20[95] with affinity $<$ µM to the PDSP targets using the RDKit Morgan algorithm (an ECFP implementation) as well as by a preliminary version of E3FP (Table S4). We likewise fingerprinted the ZINC15 compounds using both ECFP and E3FP. We queried the search compounds using SEA against a discrete sets of ligands from < 10 nM affinity (strong binders) to $<$ µM affinity (weak binders) to each target, in log-order bins, using both ECFP and E3FP independently. We filtered the resulting predictions down to those with a strong SEA-E3FP *p-value* $< 1 \times 10^{-25}$ and

$\leq$ nM affinity to the target, where the SEA-ECFP *p-value* exceeded 0.1 (*i.e.*, there was no significant SEA-ECFP prediction) in the same log-order affinity bin. From this set of compound-target pairs, we manually selected eight for experimental testing.

Experimental Assays of Compound-Target Binding Pairs

Radioligand binding and functional assays were performed as previously described[71,96,97]. Detailed experimental protocols and curve fitting procedures are available on the NIMH PDSP website at: `https://pdspdb.unc.edu/pdspWeb/content/PDSP%20Protocols%20II%202013-03-28.pdf`.

Ligand efficiencies were calculated using the expression

$$LE = -RT(\ln K_i)/N_{\text{heavy}} \approx -0.596 \ln K_i/N_{\text{heavy}}$$

where R is the ideal gas constant, T is the experimental temperature in Kelvin, and N_{heavy} is the number of heavy atoms in the molecule[98]. The ligand efficiency is expressed in units of kcal/mol/heavy atom.

Source Code

Code for generating E3FP fingerprints is available at `https://github.com/keiserlab/e3fp/tree/1.0` under the GNU Lesser General Public License version 3.0 (LGPLv3) license. All code necessary to reproduce this work is available at `https://github.com/keiserlab/e3fp-paper/tree/1.0` under the GNU LGPLv3 license.

2.6 Supporting Information

Supporting figures and tables include an enlarged **Fig. 2.1**c, parameter optimization and cross-validation results, references for highlighted molecule pairs in **Fig. 2.3**, descriptions of compounds used in experiments, and all experimental results. This material is available free of charge via the Internet at doi:10.1021/acs.jmedchem.7b00696.

Acknowledgments

This material is based upon work supported by a Paul G. Allen Family Foundation Distinguished Investigator Award (to MJK), a New Frontier Research Award from the Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation (to MJK), and the National Science Foundation Graduate Research Fellowship Program under Grant No. 1650113 (to SDA and ELC). ELC is a Howard Hughes Medical Institute Gilliam Fellow. K_i determinations and agonist and antagonist functional data was generously provided by the National Institute of Mental Health's Psychoactive Drug Screening Program, Contract # HHSN-271-2013-00017-C (NIMH PDSP). The NIMH PDSP is Directed by Bryan L. Roth MD, PhD at the University of North Carolina at Chapel Hill and Project Officer Jamie Driscoll at NIMH, Bethesda MD, USA. We thank Teague Stirling, Michael Mysinger, Cristina Melero, John Irwin, William DeGrado, and Brian Shoichet for discussions and technical support.

Abbreviations Used

AUPRC, AUC of the Precision-Recall Curve; AUROC, AUC of the Receiver Operating Characteristic Curve; E3FP, Extended Three-Dimensional FingerPrint; ECFP, Extended Connectivity FingerPrint; NB, Naive Bayes Classifier; NN, Artificial Neural Network; PRC,

Precision-Recall Curve; RF, Random Forest; ROC, Receiver Operating Characteristic Curve; SEA, Similarity Ensemble Approach; SVM, Support Vector Machine; TC, Tanimoto coefficient

References

1. Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *Journal of Medicinal Chemistry* **40,** 1219–1229. doi:10.1021/jm960352+ (Apr. 1997) (cit. on p. 41).

2. Nicholls, A., McGaughey, G. B., Sheridan, R. P., Good, A. C., Warren, G., Mathieu, M., Muchmore, S. W., Brown, S. P., Grant, J. A., Haigh, J. A., Nevins, N., Jain, A. N. & Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *Journal of Medicinal Chemistry* **53,** 3862–3886. doi:10.1021/jm900818s (May 2010) (cit. on pp. 41, 42, 61, 63).

3. Sheridan, R. P. & Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discovery Today* **7,** 903–911. doi:10.1016/S1359-6446(02)02411-X (Sept. 2002) (cit. on pp. 41, 42).

4. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. & Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nature Biotechnology* **25,** 197–206. ISSN: 1087-0156. doi:10.1038/nbt1284 (Feb. 2007) (cit. on pp. 41, 47, 59, 63, 68).

5. Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., Whaley, R., Glennon, R. A., Hert, J., Thomas, K. L. H., Edwards, D. D., Shoichet, B. K. & Roth, B. L. Predicting New Molecular Targets for Known Drugs. *Nature* **462,** 175–181. ISSN: 1476-4687. doi:10.1038/nature08506 (Nov. 2009) (cit. on pp. 41, 47, 59).

6. Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I. & Shoichet, B. K. Quantifying the Relationships among Drug Classes. *Journal of Chemical Information and Modeling* **48,** 755–765. doi:10.1021/ci8000259 (Apr. 2008) (cit. on pp. 41, 46, 60).

7. Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **57,** 3186–3204. doi:10.1021/jm401411z (Apr. 2014) (cit. on pp. 41, 60, 61).

8. Keiser, M. J., Irwin, J. J. & Shoichet, B. K. The Chemical Basis of Pharmacology. *Biochemistry* **49,** 10267–10276. doi:10.1021/bi101540g (Dec. 2010) (cit. on p. 41).

9. Zhang, H. The Optimality of Naive Bayes. *Proc. 17th. Fla. Artif. Intell. Res. Soc.,* 562–567 (Dec. 2004) (cit. on pp. 41, 59).

10. Chen, B., Harrison, R. F., Papadatos, G., Willett, P., Wood, D. J., Lewell, X. Q., Greenidge, P. & Stiefl, N. Evaluation of Machine-Learning Methods for Ligand-Based Virtual Screening. *Journal of Computer-Aided Molecular Design* **21,** 53–62. doi:10.1007/s10822-006-9096-5 (Mar. 2007) (cit. on p. 41).

11. Bender, A., Mussa, H. Y., Glen, R. C. & Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *Journal of Chemical Information and Computer Sciences* **44,** 170–178. doi:10.1021/ci034207y (Feb. 2004) (cit. on p. 41).

12. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. ISSN: 08856125 (2001) (cit. on p. 41).

13. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. & Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43,** 1947–1958. doi:10.1021/ci034160g (Dec. 2003) (cit. on p. 41).

14. Cortes, C. & Vapnik, V. Support-Vector Networks. *Machine Learning* **20,** 273–297. ISSN: 0885-6125. doi:10.1007/BF00994018 (Sept. 1995) (cit. on pp. 41, 59).

15. Franke, L., Byvatov, E., Werz, O., Steinhilber, D., Schneider, P. & Schneider, G. Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *Jour-*

nal of Medicinal Chemistry **48,** 6997–7004. doi:10.1021/jm050619h (Nov. 2005) (cit. on p. 41).

16. Dahl, G. E., Jaitly, N. & Salakhutdinov, R. Multi-Task Neural Networks for QSAR Predictions. *arXiv:1406.1231. arXiv.org e-Print archive.* (Apr. 2014) (cit. on p. 41).

17. Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D. & Pande, V. Massively Multitask Networks for Drug Discovery. *arXiv* (2015) (cit. on p. 41).

18. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *Journal of Computer-Aided Molecular Design* **30,** 595–608. doi:10.1007/s10822-016-9938-8 (Aug. 2016) (cit. on p. 41).

19. Baskin, I. I., Winkler, D. & Tetko, I. V. A Renaissance of Neural Networks in Drug Discovery. *Expert Opin Drug Discov* **11,** 785–795. doi:10.1080/17460441.2016.1201262 (Aug. 2016) (cit. on p. 41).

20. Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H. & Hochreiter, S. Deep Learning as an Opportunity in Virtual Screening. *Advances in neural information processing systems* **27** (2014) (cit. on p. 41).

21. Czodrowski, P. & Bolick, W.-G. OCEAN: Optimized Cross rEActivity estimatioN. *Journal of Chemical Information and Modeling* **56,** 2013–2023. doi:10.1021/acs.jcim.6b00067 (Oct. 2016) (cit. on p. 41).

22. Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *Journal of Medicinal Chemistry* **50,** 74–82. doi:10.1021/jm0603365 (Jan. 2007) (cit. on p. 41).

23. Rush, T. S., Grant, J. A., Mosyak, L. & Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *Journal of Medicinal Chemistry* **48,** 1489–1495. doi:10.1021/jm040163o (Mar. 2005) (cit. on p. 41).

24. Haque, I. S. & Pande, V. S. SCISSORS: A Linear-Algebraical Technique to Rapidly Approximate Chemical Similarities. *Journal of Chemical Information and Modeling* **50,** 1075–1088. doi:10.1021/ci1000136 (June 2010) (cit. on p. 42).

25. Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R. & Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Research* **42,** D1083–90. doi:10.1093/nar/gkt1031 (Jan. 2014) (cit. on pp. 42, 63).

26. Sun, J., Jeliazkova, N., Chupakin, V., Golib-Dzib, J.-F., Engkvist, O., Carlsson, L., Wegner, J., Ceulemans, H., Georgiev, I., Jeliazkov, V., Kochev, N., Ashby, T. J. & Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J Cheminform* **9,** 17. doi:10.1186/s13321-017-0203-5 (Mar. 2017) (cit. on p. 42).

27. Jenkins, J. L., Glick, M. & Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *Journal of Medicinal Chemistry* **47,** 6144–6159. doi:10.1021/jm049654z (Dec. 2004) (cit. on p. 42).

28. Jenkins, J. L. in *Scaffold Hopping in Medicinal Chemistry* (ed Brown, N.) 155–174 (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, Dec. 2013). ISBN: 978-3-527-33364-6. doi:10.1002/9783527665143.ch10 (cit. on p. 42).

29. Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C. & Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *Journal of Medicinal Chemistry* **42,** 3251–3264. doi:10.1021/jm9806998 (Aug. 1999) (cit. on p. 42).

30. Pickett, S. D., Mason, J. S. & McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *Journal of Chemical Information and Computer Sciences* **36,** 1214–1223. ISSN: 0095-2338. doi:10.1021/ci960039g (Jan. 1996) (cit. on p. 42).

31. Sterling, T. & Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **55,** 2324–2337. doi:10.1021/acs.jcim.5b00559 (Nov. 2015) (cit. on pp. 42, 63, 72).

32. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50,** 742–754. doi:10.1021/ci100050t (May 2010) (cit. on pp. 43, 51, 66).

33. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-a Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **5,** 107–113. ISSN: 0021-9576. doi:10.1021/c160017a018 (May 1965) (cit. on p. 43).

34. Barelier, S., Sterling, T., O'Meara, M. J. & Shoichet, B. K. The Recognition of Identical Ligands by Unrelated Proteins. *ACS chemical biology* **10,** 2772–2784. doi:10.1021/acschembio.5b00683 (Dec. 2015) (cit. on p. 43).

35. Ebejer, J.-P., Morris, G. M. & Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *Journal of Chemical Information and Modeling* **52,** 1146–1158. doi:10.1021/ci2004658 (May 2012) (cit. on pp. 45, 64).

36. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *Journal of Chemical Information and Computer Sciences* **44,** 1177–1185. doi:10.1021/ci034231b (June 2004) (cit. on pp. 46, 47).

37. DeGraw, A. J., Keiser, M. J., Ochocki, J. D., Shoichet, B. K. & Distefano, M. D. Prediction and Evaluation of Protein Farnesyltransferase Inhibition by Commercial

Drugs. *Journal of Medicinal Chemistry* **53,** 2464–2471. doi:10.1021/jm901613f (Mar. 2010) (cit. on p. 47).

38. Yee, S. W., Lin, L., Merski, M., Keiser, M. J., Gupta, A., Zhang, Y., Chien, H.-C., Shoichet, B. K. & Giacomini, K. M. Prediction and Validation of Enzyme and Transporter Off-Targets for Metformin. *Journal of pharmacokinetics and pharmacodynamics* **42,** 463–475. doi:10.1007/s10928-015-9436-y (Oct. 2015) (cit. on p. 47).

39. Laggner, C., Kokel, D., Setola, V., Tolia, A., Lin, H., Irwin, J. J., Keiser, M. J., Cheung, C. Y. J., Minor, D. L., Roth, B. L., Peterson, R. T. & Shoichet, B. K. Chemical Informatics and Target Identification in a Zebrafish Phenotypic Screen. *Nature Chemical Biology* **8,** 144–146. doi:10.1038/nchembio.732 (Dec. 2011) (cit. on p. 47).

40. Lemieux, G. A., Keiser, M. J., Sassano, M. F., Laggner, C., Mayer, F., Bainton, R. J., Werb, Z., Roth, B. L., Shoichet, B. K. & Ashrafi, K. In Silico Molecular Comparisons of C. Elegans and Mammalian Pharmacology Identify Distinct Targets That Regulate Feeding. *PLoS biology* **11,** e1001712. doi:10.1371/journal.pbio.1001712 (Nov. 2013) (cit. on p. 47).

41. Bruni, G., Rennekamp, A. J., Velenich, A., McCarroll, M., Gendelev, L., Fertsch, E., Taylor, J., Lakhani, P., Lensen, D., Evron, T., Lorello, P. J., Huang, X.-P., Kolczewski, S., Carey, G., Caldarone, B. J., Prinssen, E., Roth, B. L., Keiser, M. J., Peterson, R. T. & Kokel, D. Zebrafish Behavioral Profiling Identifies Multitarget Antipsychotic-like Compounds. *Nature Chemical Biology* **12,** 559–566. doi:10.1038/nchembio.2097 (July 2016) (cit. on p. 47).

42. Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Côté, S., Shoichet, B. K. & Urban, L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature* **486,** 361–367. ISSN: 0028-08361476-4687. doi:10.1038/nature11159 (June 2012) (cit. on p. 47).

43. Lorberbaum, T., Nasir, M., Keiser, M. J., Vilar, S., Hripcsak, G. & Tatonetti, N. P. Systems Pharmacology Augments Drug Safety Surveillance. *Clinical Pharmacology & Therapeutics* **97,** 151–158. doi:10.1002/cpt.2 (Feb. 2015) (cit. on p. 47).

44. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **10,** e0118432. doi:10.1371/journal.pone.0118432 (Mar. 2015) (cit. on p. 49).

45. Snoek, J., Larochelle, H. & Adams, R. Practical Bayesian Optimization of Machine Learning Algorithms. *NIPS* (2012) (cit. on pp. 49, 70).

46. Froestl, W., Mickel, S. J., Hall, R. G., von Sprecher, G., Strub, D., Baumann, P. A., Brugger, F., Gentsch, C., Jaekel, J. & Olpe, H. R. Phosphinic Acid Analogues of GABA. 1. New Potent and Selective GABAB Agonists. *Journal of Medicinal Chemistry* **38,** 3297–3312 (Aug. 1995) (cit. on p. 54).

47. Moormann, A. E., Metz, S., Toth, M. V., Moore, W. M., Jerome, G., Kornmeier, C., Manning, P., Hansen, D. W., Pitzele, B. S. & Webber, R. K. Selective Heterocyclic Amidine Inhibitors of Human Inducible Nitric Oxide Synthase. *Bioorganic & Medicinal Chemistry Letters* **11,** 2651–2653 (Oct. 2001) (cit. on p. 54).

48. Shankaran, K., Donnelly, K. L., Shah, S. K., Guthikonda, R. N., MacCoss, M., Humes, J. L., Pacholok, S. G., Grant, S. K., Kelly, T. M. & Wong, K. K. Evaluation of Pyrrolidin-2-Imines and 1,3-Thiazolidin-2-Imines as Inhibitors of Nitric Oxide Synthase. *Bioorganic & Medicinal Chemistry Letters* **14,** 4539–4544. doi:10.1016/j.bmcl.2004.06.033 (Sept. 2004) (cit. on p. 54).

49. Ponticello, G. S., Freedman, M. B., Habecker, C. N., Lyle, P. A., Schwam, H., Varga, S. L., Christy, M. E., Randall, W. C. & Baldwin, J. J. Thienothiopyran-2-Sulfonamides: A Novel Class of Water-Soluble Carbonic Anhydrase Inhibitors. *Journal of Medicinal Chemistry* **30,** 591–597 (Apr. 1987) (cit. on p. 54).

50. Shankaran, K., Donnelly, K. L., Shah, S. K., Caldwell, C. G., Chen, P., Hagmann, W. K., Maccoss, M., Humes, J. L., Pacholok, S. G., Kelly, T. M., Grant, S. K. & Wong, K. K. Synthesis of Analogs of (1,4)-3- and 5-Imino Oxazepane, Thiazepane, and Diazepane as Inhibitors of Nitric Oxide Synthases. *Bioorganic & Medicinal Chemistry Letters* **14,** 5907–5911. doi:10.1016/j.bmcl.2004.09.019 (Dec. 2004) (cit. on p. 54).

51. Moore, W. M., Webber, R. K., Fok, K. F., Jerome, G. M., Connor, J. R., Manning, P. T., Wyatt, P. S., Misko, T. P., Tjoeng, F. S. & Currie, M. G. 2-Iminopiperidine and Other 2-Iminoazaheterocycles as Potent Inhibitors of Human Nitric Oxide Synthase Isoforms. *Journal of Medicinal Chemistry* **39,** 669–672. doi:10.1021/jm950766n (Feb. 1996) (cit. on p. 54).

52. Monn, J. A., Massey, S. M., Valli, M. J., Henry, S. S., Stephenson, G. A., Bures, M., Hérin, M., Catlow, J., Giera, D., Wright, R. A., Johnson, B. G., Andis, S. L., Kingston, A. & Schoepp, D. D. Synthesis and Metabotropic Glutamate Receptor Activity of S-Oxidized Variants of (-)-4-Amino-2-Thiabicyclo-[3.1.0]Hexane-4,6-Dicarboxylate: Identification of Potent, Selective, and Orally Bioavailable Agonists for mGlu2/3 Receptors. *Journal of Medicinal Chemistry* **50,** 233–240. doi:10.1021/jm060917u (Jan. 2007) (cit. on p. 54).

53. Nersesian, D. L., Black, L. A., Miller, T. R., Vortherms, T. A., Esbenshade, T. A., Hancock, A. A. & Cowart, M. D. In Vitro SAR of Pyrrolidine-Containing Histamine H3 Receptor Antagonists: Trends across Multiple Chemical Series. *Bioorganic & Medicinal Chemistry Letters* **18,** 355–359. doi:10.1016/j.bmcl.2007.10.067 (Jan. 2008) (cit. on p. 54).

54. Rogers, G. A., Parsons, S. M., Anderson, D. C., Nilsson, L. M., Bahr, B. A., Kornreich, W. D., Kaufman, R., Jacobs, R. S. & Kirtman, B. Synthesis, in Vitro Acetylcholine-Storage-Blocking Activities, and Biological Properties of Derivatives and Analogues

of Trans-2-(4-Phenylpiperidino)Cyclohexanol (Vesamicol). *Journal of Medicinal Chemistry* **32,** 1217–1230 (June 1989) (cit. on p. 55).

55. Röver, S., Wichmann, J., Jenck, F., Adam, G. & Cesura, A. M. ORL1 Receptor Ligands: Structure-Activity Relationships of 8-Cycloalkyl-1-Phenyl-1,3,8-Triaza-Spiro[4.5]Decan-4-Ones. *Bioorganic & Medicinal Chemistry Letters* **10,** 831–834 (Apr. 2000) (cit. on p. 55).

56. Kaltenbach, R. F., Nugiel, D. A., Lam, P. Y., Klabe, R. M. & Seitz, S. P. Stereoisomers of Cyclic Urea HIV-1 Protease Inhibitors: Synthesis and Binding Affinities. *Journal of Medicinal Chemistry* **41,** 5113–5117. doi:10.1021/jm980255b (Dec. 1998) (cit. on p. 55).

57. Lam, P. Y., Ru, Y., Jadhav, P. K., Aldrich, P. E., DeLucca, G. V., Eyermann, C. J., Chang, C. H., Emmett, G., Holler, E. R., Daneker, W. F., Li, L., Confalone, P. N., McHugh, R. J., Han, Q., Li, R., Markwalder, J. A., Seitz, S. P., Sharpe, T. R., Bacheler, L. T., Rayner, M. M., Klabe, R. M., Shum, L., Winslow, D. L., Kornhauser, D. M. & Hodge, C. N. Cyclic HIV Protease Inhibitors: Synthesis, Conformational Analysis, P2/P2' Structure-Activity Relationship, and Molecular Recognition of Cyclic Ureas. *Journal of Medicinal Chemistry* **39,** 3514–3525. doi:10.1021/jm9602571 (Aug. 1996) (cit. on p. 55).

58. Xu, B., Feng, Y., Cheng, H., Song, Y., Lv, B., Wu, Y., Wang, C., Li, S., Xu, M., Du, J., Peng, K., Dong, J., Zhang, W., Zhang, T., Zhu, L., Ding, H., Sheng, Z., Welihinda, A., Roberge, J. Y., Seed, B. & Chen, Y. C-Aryl Glucosides Substituted at the 4'-Position as Potent and Selective Renal Sodium-Dependent Glucose Co-Transporter 2 (SGLT2) Inhibitors for the Treatment of Type 2 Diabetes. *Bioorganic & Medicinal Chemistry Letters* **21,** 4465–4470. doi:10.1016/j.bmcl.2011.06.032 (Aug. 2011) (cit. on p. 55).

59. Xu, G., Lv, B., Roberge, J. Y., Xu, B., Du, J., Dong, J., Chen, Y., Peng, K., Zhang, L.,

Tang, X., Feng, Y., Xu, M., Fu, W., Zhang, W., Zhu, L., Deng, Z., Sheng, Z., Welihinda, A. & Sun, X. Design, Synthesis, and Biological Evaluation of Deuterated C-Aryl Glycoside as a Potent and Long-Acting Renal Sodium-Dependent Glucose Cotransporter 2 Inhibitor for the Treatment of Type 2 Diabetes. *Journal of Medicinal Chemistry* **57,** 1236–1251. doi:10.1021/jm401780b (Feb. 2014) (cit. on p. 55).

60. Horii, S., Fukase, H., Matsuo, T., Kameda, Y., Asano, N. & Matsui, K. Synthesis and Alpha-D-Glucosidase Inhibitory Activity of N-Substituted Valiolamine Derivatives as Potential Oral Antidiabetic Agents. *Journal of Medicinal Chemistry* **29,** 1038–1046 (June 1986) (cit. on p. 55).

61. Gao, L.-J., Waelbroeck, M., Hofman, S., Van Haver, D., Milanesio, M., Viterbo, D. & De Clercq, P. J. Synthesis and Affinity Studies of Himbacine Derived Muscarinic Receptor Antagonists. *Bioorganic & Medicinal Chemistry Letters* **12,** 1909–1912 (Aug. 2002) (cit. on p. 55).

62. de Costa, B. R., Rice, K. C., Bowen, W. D., Thurkauf, A., Rothman, R. B., Band, L., Jacobson, A. E., Radesca, L., Contreras, P. C. & Gray, N. M. Synthesis and Evaluation of N-Substituted Cis-N-Methyl-2-(1-Pyrrolidinyl)Cyclohexylamines as High Affinity Sigma Receptor Ligands. Identification of a New Class of Highly Potent and Selective Sigma Receptor Probes. *Journal of Medicinal Chemistry* **33,** 3100–3110 (Nov. 1990) (cit. on p. 56).

63. Besnard, J., Ruda, G. F., Setola, V., Abecassis, K., Rodriguiz, R. M., Huang, X.-P., Norval, S., Sassano, M. F., Shin, A. I., Webster, L. A., Simeons, F. R. C., Stojanovski, L., Prat, A., Seidah, N. G., Constam, D. B., Bickerton, G. R., Read, K. D., Wetsel, W. C., Gilbert, I. H., Roth, B. L. & Hopkins, A. L. Automated Design of Ligands to Polypharmacological Profiles. *Nature* **492,** 215–220. doi:10.1038/nature11691 (Dec. 2012) (cit. on pp. 56, 72).

64. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New

Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Research* **45,** D353–D361. doi:10.1093/nar/gkw1092 (Jan. 2017) (cit. on p. 56).

65. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Research* **44,** D457–62. doi:10.1093/nar/gkv1070 (Jan. 2016) (cit. on p. 56).

66. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30 (Jan. 2000) (cit. on p. 56).

67. Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C. & Reynolds, C. H. The Role of Ligand Efficiency Metrics in Drug Discovery. *Nature Reviews Drug Discovery* **13,** 105–121. doi:10.1038/nrd4163 (2014) (cit. on p. 58).

68. Schlicker, E., Werner, U., Hamon, M., Gozlan, H., Nickel, B., Szelenyi, I. & Göthert, M. Anpirtoline, a Novel, Highly Potent 5-HT1B Receptor Agonist with Antinociceptive/Antidepressant-like Actions in Rodents. *British Journal of Pharmacology* **105,** 732–738 (Mar. 1992) (cit. on p. 58).

69. Metzenauer, P., Barnes, N. M., Costall, B., Gozlan, H., Hamon, M., Kelly, M. E., Murphy, D. A. & Naylor, R. J. Anxiolytic-like Actions of Anpirtoline in a Mouse Light-Dark Aversion Paradigm. *Neuroreport* **3,** 527–529 (June 1992) (cit. on p. 58).

70. Moore, P. A. Local Anesthesia and Narcotic Drug Interaction in Pediatric Dentistry. *Anesthesia progress* **35,** 17 (Feb. 1988) (cit. on p. 58).

71. Barnea, G., Strapps, W., Herrada, G., Berman, Y., Ong, J., Kloss, B., Axel, R. & Lee, K. J. The Genetic Design of Signaling Cascades to Record Receptor Activation. *Proc Natl Acad Sci USA* **105,** 64–69. ISSN: 1091-6490. doi:10.1073/pnas.0710487105 (Jan. 2008) (cit. on pp. 58, 73).

72. Kroeze, W. K., Sassano, M. F., Huang, X.-P., Lansu, K., McCorvy, J. D., Giguère, P. M., Sciaky, N. & Roth, B. L. PRESTO-Tango as an Open-Source Resource for Interrogation of the Druggable Human GPCRome. *Nature Structural & Molecular Biology* **22,** 362–

369. doi:10.1038/nsmb.3014 (May 2015) (cit. on p. 58).

73. Gentry, P. R., Kokubo, M., Bridges, T. M., Cho, H. P., Smith, E., Chase, P., Hodder, P. S., Utley, T. J., Rajapakse, A., Byers, F., Niswender, C. M., Morrison, R. D., Daniels, J. S., Wood, M. R., Conn, P. J. & Lindsley, C. W. Discovery, Synthesis and Characterization of a Highly Muscarinic Acetylcholine Receptor (mAChR)-Selective M5-Orthosteric Antagonist, VU0488130 (ML381): A Novel Molecular Probe. *ChemMedChem* **9,** 1677–1682. doi:10.1002/cmdc.201402051 (Aug. 2014) (cit. on pp. 58, 61).

74. Cleves, A. E. & Jain, A. N. Effects of Inductive Bias on Computational Evaluations of Ligand-Based Modeling and on Drug Discovery. *Journal of Computer-Aided Molecular Design* **22,** 147–159. doi:10.1007/s10822-007-9150-y (Apr. 2008) (cit. on p. 61).

75. Fink-Jensen, A., Fedorova, I., Wörtwein, G., Woldbye, D. P. D., Rasmussen, T., Thomsen, M., Bolwig, T. G., Knitowski, K. M., McKinzie, D. L., Yamada, M., Wess, J. & Basile, A. Role for M5 Muscarinic Acetylcholine Receptors in Cocaine Addiction. *Journal of Neuroscience Research* **74,** 91–96. doi:10.1002/jnr.10728 (Oct. 2003) (cit. on p. 61).

76. Basile, A. S., Fedorova, I., Zapata, A., Liu, X., Shippenberg, T., Duttaroy, A., Yamada, M. & Wess, J. Deletion of the M5 Muscarinic Acetylcholine Receptor Attenuates Morphine Reinforcement and Withdrawal but Not Morphine Analgesia. *Proc Natl Acad Sci USA* **99,** 11452–11457. doi:10.1073/pnas.162371899 (Aug. 2002) (cit. on p. 61).

77. Yamada, M., Lamping, K. G., Duttaroy, A., Zhang, W., Cui, Y., Bymaster, F. P., McKinzie, D. L., Felder, C. C., Deng, C. X., Faraci, F. M. & Wess, J. Cholinergic Dilation of Cerebral Blood Vessels Is Abolished in M(5) Muscarinic Acetylcholine Receptor Knockout Mice. *Proc Natl Acad Sci USA* **98,** 14096–14101. doi:10.1073/pnas.251542998 (Nov. 2001) (cit. on p. 61).

78. Chen, D. T. Alphaprodine HCl: Characteristics. *Pediatric Dentistry* **4,** 158–63 (1982) (cit. on p. 61).

79. Bird, P. US20160000815 A1 (2016) (cit. on p. 61).

80. Bird, P. US8957099 B2 (2015) (cit. on p. 61).

81. Fleisher, C. & McGough, J. Sofinicline: A Novel Nicotinic Acetylcholine Receptor Agonist in the Treatment of Attention-Deficit/Hyperactivity Disorder. *Expert opinion on investigational drugs* **23,** 1157–1163. doi:10.1517/13543784.2014.934806 (Aug. 2014) (cit. on p. 61).

82. Kent, L., Middle, F., Hawi, Z., Fitzgerald, M., Gill, M., Feehan, C. & Craddock, N. Nicotinic Acetylcholine Receptor Alpha4 Subunit Gene Polymorphism and Attention Deficit Hyperactivity Disorder. *Psychiatric Genetics* **11,** 37–40 (Mar. 2001) (cit. on p. 62).

83. Salas, R., Cook, K. D., Bassetto, L. & De Biasi, M. The Alpha3 and Beta4 Nicotinic Acetylcholine Receptor Subunits Are Necessary for Nicotine-Induced Seizures and Hypolocomotion in Mice. *Neuropharmacology* **47,** 401–407. doi:10.1016/j.neuropharm.2004.05.002 (Sept. 2004) (cit. on p. 62).

84. Zaveri, N., Jiang, F., Olsen, C., Polgar, W. & Toll, L. Novel 34 Nicotinic Acetylcholine Receptor-Selective Ligands. Discovery, Structure-Activity Studies, and Pharmacological Evaluation. *Journal of Medicinal Chemistry* **53,** 8187–8191. doi:10.1021/jm1006148 (Nov. 2010) (cit. on p. 62).

85. Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr, B* **58,** 380–388 (June 2002) (cit. on p. 63).

86. Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N. & Rognan, D. Sc-PDB: An Annotated Database of Druggable Binding Sites from the Protein Data Bank. *Journal of Chemical Information and Modeling* **46,** 717–727. doi:10.1021/ci050372x (Apr. 2006) (cit. on p. 63).

87. Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *Journal of Computational Chemistry* (Apr. 1996) (cit. on p. 63).

88. González, M. Force Fields and Molecular Dynamics Simulations. *JDN* **12,** 169–200. ISSN: 2107-7223. doi:10.1051/sfn/201112009 (2011) (cit. on p. 63).

89. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. & Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of The American Chemical Society* **114,** 10024–10035. ISSN: 0002-7863. doi:10.1021/ja00051a040 (Dec. 1992) (cit. on pp. 63, 64).

90. RDKit: Open-Source Cheminformatics (cit. on p. 64).

91. ChemAxon. Marvin (2015) (cit. on pp. 65, 70).

92. Appleby, A. MurmurHash3 (cit. on p. 66).

93. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* (2014) (cit. on p. 72).

94. Maas, A. L., Hannun, A. Y. & Ng, A. Y. *Rectifier Nonlinearities Improve Neural Network Acoustic Models* in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing* (2013) (cit. on p. 72).

95. Hersey, A. *CHEMBL20* doi:10.6019/CHEMBL.database.20 (EMBL-EBI, Feb. 2015) (cit. on p. 72).

96. Xiao, Y., Meyer, E. L., Thompson, J. M., Surin, A., Wroblewski, J. & Kellar, K. J. Rat Alpha3/Beta4 Subtype of Neuronal Nicotinic Acetylcholine Receptor Stably Expressed in a Transfected Cell Line: Pharmacology of Ligand Binding and Function. *Molecular Pharmacology* **54,** 322–333 (Aug. 1998) (cit. on p. 73).

97. Xiao, Y., Fan, H., Musachio, J. L., Wei, Z.-L., Chellappan, S. K., Kozikowski, A. P. & Kellar, K. J. Sazetidine-A, a Novel Ligand That Desensitizes Alpha4beta2 Nicotinic Acetylcholine Receptors without Activating Them. *Molecular Pharmacology* **70,** 1454–1460. doi:10.1124/mol.106.027318 (Oct. 2006) (cit. on p. 73).

98. Kuntz, I. D., Chen, K., Sharp, K. A. & Kollman, P. A. The Maximal Affinity of Ligands. *Proc Natl Acad Sci USA* **96,** 9997–10002. doi:10.1073/pnas.96.18.9997 (Aug. 1999) (cit. on p. 73).

Chapter 3

Deep learning approaches in predicting ADMET properties

3.1 Abstract

Since the early days of Lipinski's rule of five, the field of predictive ADMET (absorption, distribution, metabolism, excretion and toxicity) in medicinal chemistry has expanded in importance and has grown to include areas such as high-throughput assay development, data mining, data visualization, machine learning and structure-based modeling[1,2]. Many studies have demonstrated the role of effective application of *in silico* predictive ADMET models in accelerating the identification of small molecules with improved efficacy, safety and dose[1].

Machine learning models are now routinely used by drug discovery teams to predict properties of small molecules based off their chemical structure. The first methods were 'simple' linear models such as those used in Free-Wilson and Hammett analyses and these are still in use today due to their easy interpretability and effectiveness on small datasets[2]. Nonlinear models were later established to capture more complex relationships between structure and

activity, and such approaches include support vector machines, recursive partitioning methods (such as random forest, Cubist and XGBoost) and deep learning methods such as deep artificial neural networks (DNNs)[2,3].

Deep learning-based approaches are showing increasing promise and usefulness for ADMET prediction, fueled by increasing computational power, larger datasets generated in a standardized manner, and adaptation of image and language processing advances to chemistry[1,2]. Here, we first explore the role of deep learning in recent ADMET prediction performance advances and then discuss ongoing work to address challenges in evaluating, interpreting and implementing deep learning for molecular design.

3.2 Rise of deep learning for ADMET prediction

A 2012 Kaggle competition sponsored by Merck & Co., Inc., (NJ, USA) generated excitement around DNNs for ADMET prediction when researchers reported that simple DNNs yielded a 10% mean improvement in R^2 performance across 15 large assay datasets when compared with widely-used, prevailing random forest (RF) models[4]. R^2 is the squared Pearson correlation coefficient that ranges from 0–1 (higher is better) and measures how well the prediction matches the experimental data, usually on a left-out data subset[4]. Other researchers, including those at Vertex Inc., (MA, USA)[5], Eli Lilly & Co. (IN, USA)[6] and Bayer AG (Berlin, Germany)[7], similarly found that simple DNNs (in these cases, fully-connected DNNs) were comparable or slightly improved over prevailing models when trained on large, proprietary ADMET datasets. Despite promising results, many practitioners at the time felt that the impact of the generally modest performance improvement seen in these early DNNs did not justify the major increase in resources, computational and human, required to maintain such models.

The winning Kaggle competitor demonstrated dramatic performance improvements on some Kaggle end points using a multitask DNN that simultaneously learned all assay tasks

within a single model[4,7–9]. However, thoughtful investigation into these DNNs led to a more nuanced understanding of their performance for ADMET prediction. Xu et al. and Kearnes et al. raised concerns that multitask models derived a large part of their improved performance from 'memorizing' or 'leaking' molecules regardless of the relevance of assisting assays[2,5,7,9]. This behavior manifests advantageously when molecules are structurally similar and assays are related, but potentially worsens or does not improve predictions when molecules are similar and assays are unrelated.

In an interesting demonstration of this, Wenzel et al. trained a multitask fully-connected DNN using large Sanofi–Aventis AG liver microsome stability datasets (∼50,000 compounds) for three species and found predictive performance improved for human liver microsomes (HLM), but worsened slightly or stayed the same for rat and mouse liver microsomes (RLM and MLM, respectively)[8]. However, when overlapping molecules were disallowed between the prediction molecules for HLM and any training molecules (regardless of species), essentially no benefit was seen from multitask training. Upon addition of five additional species with small liver microsome datasets of 200–1500 compounds each to create an eight-species multitask model, Wenzel et al. saw mixed predictive performance, with macaque predictions benefiting (R^2 increased by 0.09 or 15%), but monkey predictions worsening (R^2 decreased by 0.09 or 24%). Montanari et al. explored the benefits of multitask fully-connected DNNs and concluded it was 'trial-and-error type of work'[7].

3.3 Learned featurization improves predictive performance

As enthusiasm tempered for multitask DNNs, a new twist on deep learning was emerging. Both the prevailing models (such as RF models) and the DNN models described above require a static 'fingerprint' representation of each compound. Fingerprints are vectors

of discrete numbers where each number represents the presence (zero or one) or count of a chemical fragment[2]. Duvenaud et al. proposed that instead of enumerating possible fragments into a static fingerprint, a graph convolutional DNN (GCNN) approach could be used to dynamically learn a fingerprint optimized for the most relevant chemical information, where adjacency in the vector representation encodes fragment similarity relevant to an assay end point of interest[2,5,10–14].

In a recent study, Feinberg et al. investigated prediction of 31 ADMET end points using a state-of-the-art GCNN and large, standardized assay datasets at Merck & Co., Inc.[14]. Their single-task GCNN showed strong performance improvements over RF for about a third of end points, with improvements in R^2 of 0.15–0.31 or 45%–133%. These end points include plasma protein binding (PPB), HLM and RLM, cytochrome p450 3A4 (CYP3A4) and 2D6 inhibition, hERG binding and kinetic solubility in water. Especially remarkable was the large improvement for rat and human PPB (R^2 improvements of 0.31 and 0.19, respectively). Another third of end points showed smaller improvements (R^2 improvements of 0.05–0.14) that may be meaningful. Conversely, a third of end points showed minimal or no improvement, notably including all in vivo pharmacokinetic end points. The authors saw additional performance improvements by adding a multitask approach for featurization and this improvement was most notable for hepatocyte stability, permeability and P-glycoprotein efflux. Overall, they found that their multitask GCNNs performed equivalently or better than single-task GCNNs, which, in turn, performed equivalently or better than RF models.

Using a different state-of-the-art GCNN, Yang et al. showed notable performance relative to RF on smaller proprietary industrial ADMET datasets as well as on public datasets. On four large Amgen Inc., (CA, USA), datasets (rat PPB, solubility, RLM and human PXR activation), their ensemble single-task GCNN models showed median improvement of 14% in root-mean-square error compared with RF models using fingerprints[11]. Liu et al. compared yet another single-task GCNN model approach to a Cubist approach using five different Amgen Inc. datasets[15], and found that single and multitask GCNNs showed 10% and 38%

mean improvements in R^2 performance, respectively, over Cubist models. Liu et al. found that single-task GCNNs consistently outperformed Cubist in HLM, CYP3A4 and solubility end points, while both performed equivalently for human PXR activation prediction. They applied a multitask GCNN to modeling of multiple solubility and PXR subsets and these models consistently outperformed Cubist models. These studies suggest that GCNNs are not only competitive to RF and Cubist models, but in some cases yield robust increases in predictivity.

Another study by Montanari et al. used eight Bayer AG proprietary ADMET datasets with 39,000–236,000 compounds each to compare the learned representations from GCNNs against two fingerprint-based approaches, simple fully-connected DNNs and RFs[7]. Multitask GCNNs improved predictive performance the most over prevailing RF models. With membrane affinity and PPB end points, they saw R^2 improvements of 0.26–0.28 or 65–67%. Overall, they found that deep learning approaches out-performed RF, multitask models were competitive or outperformed single-task models, and GCNNs outperformed fully-connected DNNs of the same singletask or multitask training type.

Taken together, recent work suggests models built using GCNNs can consistently deliver better or equivalent predictive performance compared with prevailing approaches (such as RF, Cubist and support vector machine), and, in some cases, demonstrate impressive gains. We highlight PPB and RLM/HLM end points as two end points that see strong improvement in multiple studies, especially with larger industrial datasets (>15,000 compounds). Multitask GCNNs appear to slightly outperform single-task GCNNs, although ensemble single-task GCNNs reported by Yang et al.[11] are putting up a solid fight for comparative performance (manuscript in preparation).

3.4 Measuring how models generalize for medicinal chemistry

The performance metrics reported in the earlier sections use datasets split into a learning subset and a testing subset using either a time-based split or a scaffold-based split[16]. These two splits are preferred over a random split, which is considered less reflective of real-world drug discovery settings and gives an unrealistically optimistic assessment of prospective performance. In time splits, earlier data is used for training, and more recent data for testing. This simulates the drug discovery setting where researchers relentlessly evolve molecules into novel chemical space over time. Recent work suggests that a scaffold split —which assigns different compound clusters or scaffolds to the training and testing sets —can be useful when temporal splits are insufficiently diverse, where there is no testing date available or where there are concerns over model generalizability. Using industrial datasets, scientists at our laboratories found models perform better on time splits than on scaffold splits of the same datasets, and both are clearly more challenging than random splits[9,16]. On the other hand, studies using Amgen Inc. (CA, USA), Novartis AG, (Basel, Switzerland), and Bayer AG (Berlin, Germany) datasets found scaffold splits to be slightly more difficult than time splits[11]. Our takeaway is that scaffold split is a good, but imperfect, estimate of time split and both are clearly more predictive of expected performance in medicinal chemistry progression than random split.

To assess the potential of GCNN models for chemical space extrapolation, Feinberg et al. recently proposed a time plus molecular weight (MW) split[14]. In this split, models are trained on earlier compounds with MW 500 Da and tested on later molecules with MW 600 Da. Under this split, GCNNs were able to maintain much of their performance on the third of assays where they performed strongly[14]. This contrasts with the poor performance of RF models. The results suggest that learned representations encoded by GCNNs may be better at extracting structure–activity relationships from smaller molecules that extrapolate

in chemical space to larger molecules.

Splits are schemes designed purely to assess predictive performance. When it comes to building models for deployment to medicinal chemistry teams, all available data are used for training, and, in practice, the predictive performance for lead optimization is typically slightly better than that reported in the studies discussed above. In general, machine learning approaches perform better with multiple experimental observations and molecular diversity. Repeated observations are useful for quantifying experimental variability in the assay and therefore the limits of predictability[1,17]. Molecular diversity in the training data allows models to generalize to a wide range of molecular scaffolds (global diversity) as well as learn nuances from smaller functional group perturbations (localized diversity). A continued discussion between modelers and chemists can lead to selection of molecules that fill in model gaps to ultimately improve predictions. We next discuss methods that can help direct such discussions.

3.5 Interpretability, error & the use of deep learning models

As DNN models increasingly are applied in drug discovery, users will want information on a model's underlying molecular assumptions, uncertainty for individual predictions and chemical domain validity. Deep learning methods currently fall short of regression and recursive partitioning approaches (such as RF and Cubist) in these areas, although addressing these shortcomings are the focus of active, burgeoning work[2,3].

Interpretability is important in allowing scientists to 'sanity check' structure–activity relationships identified by a model and identify spurious correlations[1,2] which restrict rather than improve drug discovery. Current interpretative approaches can be broken down into three types: quantifying prediction changes at a specific heavy atom from *in silico* addition

or deletion of a fragment to the atom position, quantifying the effect of a particular input feature based on a DNN's internal learned weight matrix and building separate explanation models that approximate a complex model that is not interpretable (such as Shapley additive explanations)[2,3,8,14,18]. These approaches highlight substructures on a molecule that contribute strongly to a model's prediction and enable both human review and further molecular design. We expect this area to develop significantly in the next couple years.

In addition to interpretability, users can benefit from reliable estimates for when an individual prediction is a confident one or is beyond the scope of the model's underlying assumptions. Hirschfeld et al. recently categorized error estimation approaches for deep learning methods as falling into four broad categories: using ensembles of models generated with multiple parameter or training sets, creating DNN models that directly estimate a standard deviation along with a predicted value, calculating uncertainty based on distance from closely related training molecules and combining DNN outputs with methods like RF that have established approaches to estimate uncertainty[19].

It is important to separate the concepts of errors intrinsic to assays themselves (aleatoric error) from errors caused by insufficient data and limitations of machine learning models (epistemic uncertainty) as each would result in a different approach to fix the error (such as alternate assays for aleatoric error or increased sampling of chemical space for epistemic uncertainty). Currently, there is little consensus on the overall optimal approach for estimating uncertainty for DNNs, but methods that append a RF model to a GCNN do appear to have a slight edge and error prediction performance does appear to be task dependent[19]. Emerging DNNs known as fully Bayesian networks[19], whose parameters are treated as distributions, also hold promise. As methodologies mature for error estimation, greater confidence may be applied to certain model predictions and molecular spaces to be treated with concern can be highlighted.

As an additional note, while DNNs are considered state-of-the-art for certain tasks, they

require a lot of data and computational resources. For low data regimes, traditional methods such as RF or even regression may perform as well or better than DNN approaches and should not be discarded out-of-hand, especially considering that these models have established techniques for interpretability and error estimation[1]. As the field continues to change, the goal for predictive modelers and drug discovery groups will be to continuously review trade-offs between methods.

3.6 Future perspective

We believe that GCNNs will play a major role in ADMET prediction models for medicinal chemistry and their predictive ability has already surpassed RF and fully connected DNNs for a subset of end points. So where will the next improvements in ADMET predictions come from? We suggest three areas.

The first is increased incorporation of mechanistic understanding and biophysical information that DNNs must currently attempt to infer. The model interpretability tools discussed above provide insights that can be matched to mechanistic understanding[1]. Biophysical information can include quantum chemical properties and 3D properties such as molecular flexibility, conformational energies and biomolecular interactions (ion channels and CYPs, for instance). For more complex biological end points, such as those from in vivo pharmacokinetic studies, incorporating modeling of biological processes may help[1].

The second is creative algorithmic advances to improve generalization and reduce reliance on dataset size. Currently, deep learning models may be limited by the range of the chemical diversity, data availability and bias inherent to an institution's focus. One approach from the vision and language realms called 'pretraining' allow a DNN model trained once using very large datasets to be successful in new domains with very small datasets[1,2]. While this approach has so far yielded mixed results for chemistry, it is still early days. It also remains to be seen how efforts by researchers to share and combine data from public and private

realms can overcome deficiencies in chemistry coverage[4].

Finally, ADMET prediction models can often be better integrated into the iterative medicinal chemistry design-make-test cycle. We note that in our practice, models considered poorly predictive (e.g., hERG models with time split R^2 of $\sim$0.3) are still useful in prioritizing molecules, especially in the context of large numbers of molecules and categorical predictions[20]. Nevertheless, methods for highlighting gaps in prediction model 'knowledge' can facilitate design of molecules to explore deficiencies. DNNs using continuous representations that allow for generative creation of new molecules with desired properties may further contribute to efficient lead evolution[1]. We believe a close coupling of graph convolutional ADMET predictors and medicinal chemistry collaboration will improve the speed and reliability of designing new and improved therapeutics.

3.7 Author contributions

All authors contributed to writing of this article.

3.8 Acknowledgments

The authors thank JC Alvarez, KV Chuang, MJ Keiser, E Joshi and R Sheridan for helpful comments. We unfortunately had to omit many good references given article format limitations.

3.9 Financial & competing interests disclosure

EL Cáceres is supported under the National Science Foundation Graduate Research Fellowship Program under grant no. 1650113. The authors are currently or previously employed

by Merck & Co., Inc. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

3.10 Open access

References

1. Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., Fisher, J., Jansen, J. M., Duca, J. S., Rush, T. S., Zentgraf, M., Hill, J. E., Krutoholow, E., Kohler, M., Blaney, J., Funatsu, K., Luebkemann, C. & Schneider, G. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* **19,** 353–364. doi:10.1038/s41573-019-0050-3 (Dec. 2019) (cit. on pp. 91, 92, 97, 99, 100).

2. Chuang, K. V., Gunsalus, L. M. & Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *Journal of Medicinal Chemistry* **63,** 8705–8722. doi:10.1021/acs.jmedchem.0c00385 (May 2020) (cit. on pp. 91–94, 97–99).

3. Polishchuk, P. Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. *Journal of Chemical Information and Modeling* **57,** 2618–2639. doi:10.1021/acs.jcim.7b00274 (Oct. 2017) (cit. on pp. 92, 97, 98).

4. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling* **55,** 263–274. doi:10.1021/ci500747n (Feb. 2015) (cit. on pp. 92, 93, 100).

5. Kearnes, S., Goldman, B. & Pande, V. *Modeling Industrial ADMET Data with Multitask Networks* 2017 (cit. on pp. 92–94).

6. Zhou, Y., Cahya, S., Combs, S. A., Nicolaou, C. A., Wang, J., Desai, P. V. & Shen, J. Exploring Tunable Hyperparameters for Deep Neural Networks with Industrial ADME Data Sets. *Journal of Chemical Information and Modeling* **59,** 1005–1016. doi:10.1021/acs.jcim.8b00671 (Dec. 2018) (cit. on p. 92).

7. Montanari, F., Kuhnke, L., Laak, A. T. & Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **25,** 44.

doi:10.3390/molecules25010044 (Dec. 2019) (cit. on pp. 92, 93, 95).

8. Wenzel, J., Matter, H. & Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling* **59,** 1253–1268. doi:10.1021/acs.jcim.8b00785 (Jan. 2019) (cit. on pp. 93, 98).

9. Xu, Y., Ma, J., Liaw, A., Sheridan, R. P. & Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling* **57,** 2490–2504. doi:10.1021/acs.jcim.7b00087 (Oct. 2017) (cit. on pp. 93, 96).

10. Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. & Adams, R. P. in *Advances in Neural Information Processing Systems 28* (eds Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) 2224–2232 (Curran Associates, Inc., 2015) (cit. on p. 94).

11. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. & Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **59,** 3370–3388. doi:10.1021/acs.jcim.9b00237 (July 2019) (cit. on pp. 94–96).

12. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling* **57,** 1757–1772. doi:10.1021/acs.jcim.6b00601 (July 2017) (cit. on p. 94).

13. Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., Sun, S., Yang, J., Ramsundar, B. & Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Central Science* **4,** 1520–1530. doi:10.1021/acscentsci.8b00507 (Nov. 2018) (cit. on p. 94).

14. Feinberg, E. N., Joshi, E., Pande, V. S. & Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *Journal of Medicinal Chemistry* **63,** 8835–8848. doi:10.1021/acs.jmedchem.9b02187 (Apr. 2020) (cit. on pp. 94, 96, 98).

15. Liu, K., Sun, X., Jia, L., Ma, J., Xing, H., Wu, J., Gao, H., Sun, Y., Boulnois, F. & Fan, J. Chemi-Net: A Molecular Graph Convolutional Network for Accurate Drug Property Prediction. *International Journal of Molecular Sciences* **20,** 3389. doi:10.3390/ijms20143389 (July 2019) (cit. on p. 94).

16. Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *Journal of Chemical Information and Modeling* **53,** 783–790. doi:10.1021/ci400084k (Apr. 2013) (cit. on p. 96).

17. Sheridan, R. P., Karnachi, P., Tudor, M., Xu, Y., Liaw, A., Shah, F., Cheng, A. C., Joshi, E., Glick, M. & Alvarez, J. Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure–Activity Relationship Models? *Journal of Chemical Information and Modeling* **60,** 1969–1982. doi:10.1021/acs.jcim.9b01067 (Mar. 2020) (cit. on p. 97).

18. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *Journal of Medicinal Chemistry* **63,** 8761–8777. doi:10.1021/acs.jmedchem.9b01101 (Sept. 2019) (cit. on p. 98).

19. Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R. & Coley, C. W. *Uncertainty Quantification Using Neural Networks for Molecular Property Prediction* 2020 (cit. on p. 98).

20. Sanders, J. M., Beshore, D. C., Culberson, J. C., Fells, J. I., Imbriglio, J. E., Gunaydin, H., Haidle, A. M., Labroli, M., Mattioni, B. E., Sciammetta, N., Shipe, W. D., Sheridan, R. P., Suen, L. M., Verras, A., Walji, A., Joshi, E. M. & Bueters, T. Informing the Selection of Screening Hit Series with in Silico Absorption, Distribution,

Metabolism, Excretion, and Toxicity Profiles. *Journal of Medicinal Chemistry* **60,** 6771–6780. doi:10.1021/acs.jmedchem.6b01577 (May 2017) (cit. on p. 100).

Chapter 4

Target deconvolution for reduction of free tau in a high throughput screen

4.1 Abstract

Tau aggregation and dysfunction is a known hallmark of Alzheimer's disease and many other dementias. As such, targeting tau aggregation is an attractive candidate for therapeutics. However, the understanding of how free tau is regulated is limited. To better understand the role of tau aggregation in disease, it is important to identify cellular mechanisms which are responsible for clearing free tau and also to find tool compounds which may allow more precise inquiry. In this study, we designed a computational analysis to find mechanistic hypotheses for a recent high throughput chemical screen measuring free tau reduction. We used the Similarity Ensemble Approach to generate protein target predictions based on the molecules in the screen. We filtered protein predictions based on expression and performed hypergeometric enrichment to rank protein targets. Using our procedure, we reduced the potential search space from 2,713 protein targets to 48 targets. These targets represent hypotheses for mechanism of action which may be tested to study free tau.

4.2 Introduction

Alzheimer's disease (AD) and dementia-related diseases are marked by devastating progressive memory degradation and cognitive decline. In the US alone, AD affects 6.2 million people and in conjunction with other dementias, causes one in three senior deaths. While deaths from heart disease have decreased by 7.3% from 2000-2019, deaths from AD have increased 145%[1]. Alzheimer's disease exhibits substantial caregiver burden. Caregivers are often family members of the afflicted and approximately 11 million Americans provide 15.3 billion hours of unpaid care for Alzheimer's patients[2].

Current treatment options for AD are limited and are not effective as curative treatments. There are only five FDA-approved drugs which fit into two classes of medications for AD: cholinesterase inhibitors and N-methyl-D-aspartate (NMDA) antagonists[3]. Cholinesterase inhibitors (donepezil, rivastigmine, and galantamine), work by inhibiting the breakdown of the neurotransmitter acetylcholine by acetylcholinesterase[4]. This prolongs the presence of acetylcholine in the brain and offsets the loss of neurons which produce acetylcholine. The second class of drugs consists of the NMDA antagonist memantine. Memantine works by reducing the excitotoxic effect of excess glutamate in the brain[5]. Due to the dearth of treatment and curative options, identification of new pathways and targets for AD remains a potentially impactful research direction.

AD proteinopathy is characterized by aggregates of amyloid-beta plaques (Aβ) and tau neurofibrillary tangles (NFT), which are present in AD patients preceding clinical symptom onset[6]. Despite AD's classification as a secondary tauopathy, the tau burden of AD patients is better correlated with atrophy in the brain than Aβ[7]. Additionally, direct evidence of mutations to MAPT have been implicated in dementia[8]. This evidence suggests that targeting tau aggregation may prove useful for therapeutic interventions.

Under normal conditions, tau is a microtubule-associated phosphoprotein which helps to stabilize the microtubule. In AD and other neurodegenerative diseases, hyperphosphorylated

tau detaches from the microtubules and forms aggregates[9]. Although hyperphosphorylation is most well studied, tau is known to undergo many post-translational modifications (including ubiquitination, acetylation, and glycosylation) which have been implicated in its aggregation[10]. As free tau off of the microtubules is most available for aggregation, finding compounds which mediate its aggregation would be useful in studying dementias and the role of tau in neurodegeneration.

To better understand how free tau affects AD and other tauopathies, it is important to identify chemical tool compounds and proteins that reduce free tau *in vitro*. Our collaborators conducted a high-throughput screen for compounds associated with reduction in free tau from an inducible cell model. We sought to infer protein targets for tau reduction from the chemical similarity of screening compounds to compounds with known protein annotations. We made protein target predictions for each compound using the Similarity Ensemble Approach (SEA)[11]. We reduced our protein targets based on annotations from Uniprot and the Human Protein Atlas[12,13]. Finally, we calculated the enrichment of our targets against the background and corrected for multiple hypotheses. Our process significantly reduced the initial number of targets to a manageable set of proteins. These target hypotheses can be used to test biological hypotheses for reduction of free tau.

4.3 Results and Discussion

To better understand the mechanisms driving free tau clearance, we designed a computational pipeline to perform target deconvolution on chemical high throughput screens **Fig. 4.1**. Given 153,588 compound activities from a screen designed to identify reduction in GFP-bound free tau, we predicted a set 2,713 protein targets of interest across all compounds in the screen using SEA. Using compound hit annotations for reducing GFP tau, we removed 1,532 proteins from our enrichment list because they were not associated with any compound hits in the initial screen. Then, we mapped our ChEMBL26 protein targets to UniProt to

identify species of origin. We removed 111 non-vertebrate proteins from the enrichment set while keeping 82 proteins that did not map to a Uniprot ID and 295 proteins from non-human vertebrates. We reasoned that proteins not expressed in HEK293 cells are unlikely to be potential mechanisms of action in our screen. Therefore, we removed 318 human proteins from our enrichment list which were not expressed in HEK293. This filtration procedure reduced our enrichment set by 28%, but the set of 753 proteins was still quite large with no way to rank predictions. We ranked proteins using a hypergeometric test to calculate a p-value describing the enrichment of proteins in our hit compounds against the background of all compounds and we corrected for multiple hypotheses using the Benjamini-Hochberg correction. This final step reduced our enrichment set to 48 protein ranked targets of interest (Table 1), which represent 1.8% of our original predicted space.

The next step is to validate our procedure by selecting tool compounds from our enriched targets and to test them in vitro, with the goal of establishing a phenocopy where free tau is reduced. In the meanwhile, a preliminary literature search revealed some targets related to AD, Aβ, and tau, with two candidates currently in clinical trials:

- Tau acts on p53-binding protein MDM-2 (MDM2) resulting in deregulation of MDM2 and disruption of p53 activity and function14.
- Nuclear factor erythroid 2-related factor 2 (NRF2) is hypothesized to reduce phosphorylated tau by inducing autophagy[14].
- Protein Kinase C alpha activation is associated with reduction in soluble Aβ and also has been hypothesized to reduce tau hyperphosphorylation through GSK-3β[15].
- A Protein Kinase C inhibitor, bryostatin, is currently in phase 2 clinical trials for Alzheimer's disease[16].
- Deficiency of G protein-coupled receptor kinase 5 (GRK5), induces tau hyperphosphorylation through GSK3β activation[17].

- Disruption of the IκB kinase (IKK) complex significantly reduced tau aggregation in a CRISPR screen[18].
- ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1 (CD38), has been shown to reduce Aβ plaques in mice and is currently the target of a phase 2 clinical trial for Alzheimer's disease[19,20].
- Dual-specificity tyrosine-phosphorylation regulated kinase 1A increases Tau phosphorylation[21,22].
- Beta amyloid A4 protein (Aβ) regulates tau proteostasis[23].
- c-Jun N-terminal kinase 1 (JNK1) hyperphosphorylates tau[24].

Computational target enrichment and deconvolution of a high throughput screen on HEK293 GFP-tau/mCh-MAP2 cell lines identified known protein targets which affect tau phosphorylation and which may be involved with AD. While we tried to provide a few examples where our method may have resulted in true positives, we note that this analysis will invariably produce false positives and cannot replace compound-target testing for a phenocopy. As SEA relies on inference to a background dataset, certain predictions may be missed due to lack of data sampling or quality of annotations. Additionally, our enrichment relies on compound hits sharing mechanisms of action and that those mechanisms are unique against the background of the screen. We believe the large number of compounds selected for diversity were sufficient to ensure a suitable background, but we acknowledge there are no guarantees for the shared mechanisms of our hits despite some promising preliminary results. As we expect many of the compound hits originating from the high throughput screen to be inhibitors, our enriched proteins may not fully capture complementary effects from protein activation. An approach to complement a chemical screen could mirror research already completed on the complementarity of target predictions from CRISPRi and CRISPRa screens[25,26].

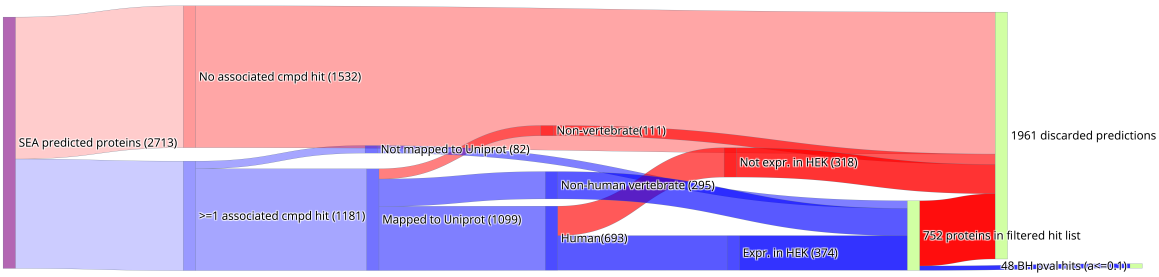

Figure 4.1: We refined the set of all SEA predicted hits across all compounds (far left) by their association with a known hit, presence in a vertebrate species, and expression in HEK293 cell lines. Our process reduced the number of putative targets from 2,713 to 752 as potential candidates for mechanism of action. We computed the hypergeometric enrichment of each target compared to the background and corrected the resulting p-values via Benjamini Hochberg multiple hypothesis correction (alpha 0.1). The final list of 48 proteins for most likely candidates represent 1.8% of the original search space. (far right).

Despite the limitations of our method, we note that this approach allowed us to reduce our search space to 1.8% of the original 2,713 protein target predictions. The method can be scaled and automated, and choices in library creation and cutoffs are easily applied equally to all targets and molecules. Data availability and completeness will affect any prediction based on reported interactions[27], but we note that our target deconvolution is not intended to state that a target does not work, it is intended to prioritize ones we have evidence for. Pragmatically, the ability to generate molecular hypotheses for biological phenotypes such as reduction in free tau is a useful tool for drug discovery and is helpful for prioritizing targets. *In silico* approaches such as our own have already demonstrated potential in uncovering interesting biological properties or targets based on chemical structure or phenotype[28–30], and we are excited for their continued utility.

Table 4.1: Ranked list of SEA predicted ChEMBL26 protein targets. Compound counts represent the number of hit or background molecules in the screen mapping to the target of interest. The p-values are Benjamini-Hochberg corrected p-values from the hypergeometric test.

Protein Description	ChEMBL ID	Hit Compounds	All Compounds	P-value
Cytochrome P450 1A2	CHEMBL3356	11	244	0.00E+00
Prenyl protein specific protease	CHEMBL3411	13	130	0.00E+00
p53-binding protein Mdm-2	CHEMBL5023	9	256	8.43E-09
Nuclear factor erythroid 2-related factor 2	CHEMBL1075094	12	558	1.33E-07
Arachidonate 5-lipoxygenase	CHEMBL5211	8	220	5.50E-07
Dual specificity tyrosine-phosphorylation-regulated kinase 1B	CHEMBL5543	11	512	5.94E-07
Mitogen-activated protein kinase; ERK1/ERK2	CHEMBL1907606	5	59	2.63E-06
Cyclin-dependent kinase 4	CHEMBL331	6	127	2.63E-06
Protein kinase C alpha	CHEMBL299	9	395	9.42E-06
CDK3/Cyclin E	CHEMBL3038471	4	35	1.31E-05
Transient receptor potential cation channel subfamily M member 8	CHEMBL3108632	6	146	1.31E-05
G protein-coupled receptor kinase 6	CHEMBL6144	9	439	3.82E-05
Thioredoxin reductase	CHEMBL2096978	5	97	5.22E-05
Glutamate [NMDA] receptor subunit epsilon 2	CHEMBL3442	4	47	7.12E-05
Cyclooxygenase	CHEMBL2095157	9	505	7.12E-05
Dual specificity protein kinase CLK3	CHEMBL4226	7	284	7.14E-05
Dual specificity protein kinase CLK2	CHEMBL4225	7	293	1.70E-04
Dual-specificity tyrosine-phosphorylation regulated kinase 1A	CHEMBL2292	10	671	1.94E-04
Ribosomal protein S6 kinase alpha 1	CHEMBL2553	5	123	1.94E-04
Platelet-derived growth factor receptor	CHEMBL2095189	11	918	5.74E-04
Aryl hydrocarbon receptor	CHEMBL3201	6	244	6.77E-04
Glutathione reductase	CHEMBL2755	5	154	1.05E-03
Dual specificity protein kinase CLK4	CHEMBL4203	9	639	1.33E-03
Serine/threonine-protein kinase SMG1	CHEMBL1795195	5	163	1.47E-03
Muscarinic acetylcholine receptor	CHEMBL2094109	5	187	1.82E-03
Phosphodiesterase 5A	CHEMBL3478	6	311	2.04E-03
G protein-coupled receptor kinase 5	CHEMBL5678	5	201	2.21E-03
Histone deacetylase 3/NCoR1	CHEMBL3038484	8	591	2.51E-03
Inhibitor of nuclear factor kappa B kinase epsilon subunit	CHEMBL3529	5	204	2.63E-03
Dual specificty protein kinase CLK1	CHEMBL4224	8	620	3.57E-03
Phosphatidylinositol 3-kinase catalytic subunit type 3	CHEMBL1075165	4	124	4.01E-03
Proto-oncogene tyrosine-protein kinase MER	CHEMBL5331	5	238	4.03E-03
Thioredoxin reductase 1, cytoplasmic	CHEMBL6035	3	62	7.05E-03
Phosphodiesterase 5A	CHEMBL4567	3	89	9.16E-03
CDK2/Cyclin A	CHEMBL3038469	6	487	9.85E-03
Kinesin-like protein 1	CHEMBL4581	9	1036	1.00E-02
Beta-adrenergic receptor kinase 1	CHEMBL3711550	6	502	1.13E-02
Histone-lysine N-methyltransferase EZH2	CHEMBL2189110	3	102	2.35E-02
ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1	CHEMBL3425388	2	29	2.75E-02
Phosphatidylinositol-5-phosphate 4-kinase type-2 gamma	CHEMBL1770034	3	110	3.67E-02
CRY2/PER2	CHEMBL4296116	3	115	3.84E-02
Phosphodiesterase 1	CHEMBL2097166	3	115	4.07E-02
Serine/threonine-protein kinase TBK1	CHEMBL5408	5	392	4.07E-02
Dual specificity tyrosine-phosphorylation-regulated kinase 1A	CHEMBL5508	4	244	4.34E-02
Beta amyloid A4 protein	CHEMBL2487	11	1590	4.91E-02
c-Jun N-terminal kinase, JNK	CHEMBL2096667	3	128	5.35E-02
Tyrosine-protein kinase ABL	CHEMBL1862	14	2343	6.93E-02
Choline acetylase	CHEMBL3945	3	133	6.93E-02

4.4 Methods

4.4.1 HEK293 GFP-tau/mCh-MAP2 cell line

Human HEK293 cell lines were designed to express a GFP-tagged wild type 0N4R tau when treated with doxycycline (Dox). To engineer cell lines that display dose-dependent control of GFP-tau coupled to the microtubules, the cells were engineered to express an *Escheria coli* dihydrofolate reductase mutant (ecDHFR)[31] fused to a MAP2c with an mCherry reporter[32]. Under untreated conditions, the ecDHFR-mCherry-MAP2c fusion will be degraded by cellular processes due to the ecDHFR's instability. However, upon treatment with trimethoprim (TMP), the ecDHFR is stabilized and ecDHFR-mCherry-MAP2c will not be degraded. MAP2c is closely related to tau (without tau's aggregation motifs) and the two proteins share microtubule binding sites[33]. The TMP-stabilized ecDHFR-mCherry-MAP2c displaces GFP-tau from the microtubules, allowing TMP-mediated control over free GFP-tau in the cytoplasm.

4.4.2 High Throughput Screens

The screening sets of compounds (n=153,588) originated from the Small Molecule Discovery Center (SMDC) at UCSF. The first set of 50k screening compounds consisted of ChemBridge "Premium" compounds. The remaining compounds were screened in a second batch and consisted of "ChemBridge ION-Kinase", "ChemBridge Gallo", and "ChemDiv Diverse" compound sets. For the primary screen, HEK293 GFP-tau/mCh-MAP2 cells were plated in PDL-coated 384-well µclear plates at 3,000 cells per well in 40 μL DMEM. The following day, 10 μL Dox/TMP solution were dispensed into plates using an EL406 liquid dispenser (BioTek) so the final concentrations of dox and TMP are 10 ng/mL and 1 μM, respectively. 100 nL of compounds were pin-transferred into plates to the final concentration of 10 µM using Biomek FXP liquid handler (Beckman Coulter).Each plate contained a vehicle con-

trol, DMSO, and 10 M FK-506 as positive control. Post 24-hour incubation, cells were then fixed and stained in 4% paraformaldehyde (Fisher Scientific) and 1 μg/mL Hoechst 33342 for 30 mins. High-throughput imaging at 20X was performed on an InCell 2500 automated microscope (GE Healthcare) and the level of GFP intensity per cell was quantified using the InCell Developer high-content image analysis software.

4.4.3 Primary screen for modulators of free GFP-tau

HEK293 GFP-tau/mCh-MAP2 cells were plated in PDL-coated 384-well µclear plates at 3,000 cells per well in 40 μL DMEM. The following day, 10 µL Dox/TMP solution were dispensed into plates using an EL406 liquid dispenser (BioTek) so the final concentrations of dox and TMP are 10 ng/mL and 1 μM, respectively. 100 nL of compounds were pin-transferred into plates to the final concentration of 10 μM using Biomek FXP liquid handler (Beckman Coulter) and incubated for 24 hr at 37°C. Cells were then fixed and stained in 4% paraformaldehyde (Fisher Scientific) and 1 μg/mL Hoechst 33342 for 30 mins. High-throughput imaging at 20X was performed on an InCell 2500 automated microscope (GE Healthcare) and the level of GFP intensity per cell was quantified using the InCell Developer high-content image analysis software.

4.4.4 Annotation of compound hits

Compound results from the primary screen were selected for activity against normalized GFP intensity corresponding to free GFP tau.

For the first screen, 145 compounds were annotated as hits if either DxA fold change or DxA b-score were less than 3 standard deviations below the mean. For this screen, toxicity of each compound was noted, but was not considered during the hit-calling process.

In the second batch screen, 1558 compounds were determined to be hits by fulfilling

criteria corresponding to activity beyond 3 standard deviations from the mean (GFP-Tau b-score > 10.13; Negative control fold change (<0.8) - normalized to only negative control; and Control normalized inhibition ($>70\%$) - normalized to both positive and negative controls). The pool of initial compounds was narrowed down to 1,199 compounds by removing toxic and proliferative compounds outside of 3 standard deviations from the mean. Cutoffs for toxic and proliferative means a compound was discarded if the cell count b-score was outside the range (-5,5) and if the negative control cell count fold change was outside of the range (0.6, 1.4). Finally, to select for the most active hits, 309 compounds were selected as final hits based on fulfilling at least two of the criteria used for activity selection. From the pool of 309 compounds, 199 compounds were selected for re-testing and 173 replicated the original active result.

For enrichment analysis, all 145 hits from the first screen, the 110 untested hits from the second screen, and the 173 validated compounds from the second screen were selected as hits.

4.4.5 SEA library preparation and annotation

We collected small molecule and protein binding data from ChEMBL26, and limited the number of protein targets to those with at least 15 binders at 10 μM or stronger[34]. We computed a statistical background for SEA using this data and computed the p-value based on an extreme value distribution and the maximum Tanimoto similarity of the prediction to the annotated compounds (MaxTc). We condensed multiple results from SEA by assigning a hit based on lowest p-val. From the condensed list of predictions, we assigned a protein hit if a target's predicted p-val for a compound was below 1e-40 or if the MaxTc of the query compound was greater than or equal to 0.40[35].

4.4.6 SEA results filtering

From the list of 2,713 condensed Similarity Ensemble Approach (SEA) predicted proteins saved for hypergeometric enrichment, we discarded 1,532 predicted targets which had no hit compounds from either screen associated with its prediction. We mapped each protein to UniProt to get species annotations and saved 82 targets which did not map to UniProt for hypergeometric enrichment. Of the 1,099 targets that mapped to UniProt, we removed 111 targets which belonged to non-vertebrate species from our enrichment list and kept 295 targets which mapped to non-human vertebrates. We removed targets from the enrichment list which were not expressed (NX <1.0) in HEK293 RNA-seq transcript data obtained from the Human Protein Atlas. Our filtration procedure removed 1,961 proteins from our final enrichment list (n=752 proteins).

4.4.7 Enrichment of protein target hits

We conducted a hypergeometric enrichment on targets associated with hit compounds compared to all compounds in the screen as background. The resulting probability for each protein describes the chances that protein would have at least the number of hit compounds associated with it given the number of all compounds associated with that protein if drawn at random. Because we had multiple p-values for comparison, we conducted a Benjamini-Hochberg multiple hypothesis correction at a permissive alpha (alpha=0.1) since we were willing to allow a slightly higher false positive rate[36]. The final list of 48 enriched proteins to search represented 1.8% of the original 2,713 SEA hits.

References

1. SCI Facts and Figures. en. *J. Spinal Cord Med.* **40,** 872–873 (Nov. 2017) (cit. on p. 107).

2. 2021 Alzheimer's disease facts and figures. en. *Alzheimers. Dement.* **17,** 327–406 (Mar. 2021) (cit. on p. 107).

3. *How is Alzheimer's disease treated?* `https://www.nia.nih.gov/health/how-alzheimers-disease-treated`. Accessed: 2021-5-26 (cit. on p. 107).

4. Joe, E. & Ringman, J. M. Cognitive symptoms of Alzheimer's disease: clinical management and prevention. en. *BMJ* **367,** l6217 (Dec. 2019) (cit. on p. 107).

5. Witt, A., Macdonald, N. & Kirkpatrick, P. Memantine hydrochloride. en. *Nat. Rev. Drug Discov.* **3,** 109–110 (Feb. 2004) (cit. on p. 107).

6. Dugger, B. N. & Dickson, D. W. Pathology of Neurodegenerative Diseases. en. *Cold Spring Harb. Perspect. Biol.* **9** (July 2017) (cit. on p. 107).

7. Whitwell, J. L., Josephs, K. A., Murray, M. E., Kantarci, K., Przybelski, S. A., Weigand, S. D., Vemuri, P., Senjem, M. L., Parisi, J. E., Knopman, D. S., Boeve, B. F., Petersen, R. C., Dickson, D. W. & Jack Jr, C. R. MRI correlates of neurofibrillary tangle pathology at autopsy: a voxel-based morphometry study. en. *Neurology* **71,** 743–749 (Sept. 2008) (cit. on p. 107).

8. Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A., Hackett, J., Adamson, J., Lincoln, S., Dickson, D., Davies, P., Petersen, R. C., Stevens, M., de Graaff, E., Wauters, E., van Baren, J., Hillebrand, M., Joosse, M., Kwon, J. M., Nowotny, P., Che, L. K., Norton, J., Morris, J. C., Reed, L. A., Trojanowski, J., Basun, H., Lannfelt, L., Neystat, M., Fahn, S., Dark, F., Tannenberg, T., Dodd, P. R., Hayward, N., Kwok, J. B., Schofield, P. R., Andreadis, A., Snowden, J., Craufurd, D., Neary, D., Owen, F., Oostra, B. A., Hardy, J., Goate, A., van Swieten, J., Mann, D., Lynch, T. & Heutink, P. Association

of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. en. *Nature* **393,** 702–705 (June 1998) (cit. on p. 107).

9. Dolan, P. J. & Johnson, G. V. W. The role of tau kinases in Alzheimer's disease. en. *Curr. Opin. Drug Discov. Devel.* **13,** 595–603 (Sept. 2010) (cit. on p. 108).

10. Arakhamia, T., Lee, C. E., Carlomagno, Y., Duong, D. M., Kundinger, S. R., Wang, K., Williams, D., DeTure, M., Dickson, D. W., Cook, C. N., Seyfried, N. T., Petrucelli, L. & Fitzpatrick, A. W. P. Posttranslational Modifications Mediate the Structural Diversity of Tauopathy Strains. en. *Cell* **180,** 633–644.e12 (Feb. 2020) (cit. on p. 108).

11. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. & Shoichet, B. K. Relating protein pharmacology by ligand chemistry. en. *Nat. Biotechnol.* **25,** 197–206 (Feb. 2007) (cit. on p. 108).

12. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L. & Ponten, F. Towards a knowledge-based Human Protein Atlas. en. *Nat. Biotechnol.* **28,** 1248–1250 (Dec. 2010) (cit. on p. 108).

13. *The Human Protein Atlas* `http://www.proteinatlas.org`. Accessed: 2021-5-12 (cit. on p. 108).

14. Jo, C., Gundemir, S., Pritchard, S., Jin, Y. N., Rahman, I. & Johnson, G. V. W. Nrf2 reduces levels of phosphorylated tau protein by inducing autophagy adaptor protein NDP52. en. *Nat. Commun.* **5,** 3496 (Mar. 2014) (cit. on p. 109).

15. Hongpaisan, J., Sun, M.-K. & Alkon, D. L. PKC ϵ activation prevents synaptic loss, Aβ elevation, and cognitive deficits in Alzheimer's disease transgenic mice. en. *J. Neurosci.* **31,** 630–643 (Jan. 2011) (cit. on p. 109).

16. *Bryostatin Treatment of Moderately Severe Alzheimer's Disease* `https://clinicaltrials.gov/ct2/show/NCT04538066?term=bryostatin&draw=2&rank=3`. Accessed: 2021-5-27 (cit. on p. 109).

17. Zhao, J., Li, X., Chen, X., Cai, Y., Wang, Y., Sun, W., Mai, H., Yang, J., Fan, W., Tang, P., Ou, M., Zhang, Y., Huang, X., Zhao, B. & Cui, L. GRK5 influences the phosphorylation of tau via GSK3β and contributes to Alzheimer's disease. en. *J. Cell. Physiol.* **234,** 10411–10420 (July 2019) (cit. on p. 109).

18. Duan, L., Hu, M., Tamm, J. A., Grinberg, Y. Y., Shen, F., Chai, Y., Xi, H., Gibilisco, L., Ravikumar, B., Gautam, V., Karran, E., Townsend, M. & Talanian, R. V. Arrayed CRISPR reveals genetic regulators of tau aggregation, autophagy and mitochondria in Alzheimer's disease model. en. *Sci. Rep.* **11,** 2879 (Feb. 2021) (cit. on p. 110).

19. Blacher, E., Dadali, T., Bespalko, A., Haupenthal, V. J., Grimm, M. O. W., Hartmann, T., Lund, F. E., Stein, R. & Levy, A. *Alzheimer's disease pathology is attenuated in a CD38-deficient mouse model* 2015 (cit. on p. 110).

20. *Study of Daratumumab in Patients With Mild to Moderate Alzheimer's Disease - Full Text View - ClinicalTrials.Gov* `https://clinicaltrials.gov/ct2/show/NCT04070378`. Accessed: 2021-5-27 (cit. on p. 110).

21. Jin, N., Yin, X., Gu, J., Zhang, X., Shi, J., Qian, W., Ji, Y., Cao, M., Gu, X., Ding, F., Iqbal, K., Gong, C.-X. & Liu, F. Truncation and Activation of Dual Specificity Tyrosine Phosphorylation-regulated Kinase 1A by Calpain I: A MOLECULAR MECHANISM LINKED TO TAU PATHOLOGY IN ALZHEIMER DISEASE*. *J. Biol. Chem.* **290,** 15219–15237 (June 2015) (cit. on p. 110).

22. Woods, Y. L., Cohen, P., Becker, W., Jakes, R., Goedert, M., Wang, X. & Proud, C. G. *The kinase DYRK phosphorylates protein-synthesis initiation factor eIF2B at Ser539 and the microtubule-associated protein tau at Thr212: potential role for DYRK as a glycogen synthase kinase 3-priming kinase* 2001 (cit. on p. 110).

23. Moore, S., Evans, L. D. B., Andersson, T., Portelius, E., Smith, J., Dias, T. B., Saurat, N., McGlade, A., Kirwan, P., Blennow, K., Hardy, J., Zetterberg, H. & Livesey, F. J. APP metabolism regulates tau proteostasis in human cerebral cortex neurons. en. *Cell*

Rep. **11,** 689–696 (May 2015) (cit. on p. 110).

24. Yoshida, H., Hastie, C. J., McLauchlan, H., Cohen, P. & Goedert, M. Phosphorylation of microtubule-associated protein tau by isoforms of c-Jun N-terminal kinase (JNK). en. *J. Neurochem.* **90,** 352–358 (July 2004) (cit. on p. 110).

25. Jost, M., Chen, Y., Gilbert, L. A., Horlbeck, M. A., Krenning, L., Menchon, G., Rai, A., Cho, M. Y., Stern, J. J., Prota, A. E., Kampmann, M., Akhmanova, A., Steinmetz, M. O., Tanenbaum, M. E. & Weissman, J. S. Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent. en. *Mol. Cell* **68,** 210–223.e6 (Oct. 2017) (cit. on p. 110).

26. Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M. & Weissman, J. S. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. en. *Cell* **159,** 647–661 (Oct. 2014) (cit. on p. 110).

27. Mestres, J., Gregori-Puigjané, E., Valverde, S. & Solé, R. V. The topology of drug–target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.* **5,** 1051–1057 (2009) (cit. on p. 111).

28. Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Côté, S., Shoichet, B. K. & Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. en. *Nature* **486,** 361–367 (June 2012) (cit. on p. 111).

29. Scheiber, J., Jenkins, J. L., Sukuru, S. C. K., Bender, A., Mikhailov, D., Milik, M., Azzaoui, K., Whitebread, S., Hamon, J., Urban, L., Glick, M. & Davies, J. W. Mapping adverse drug reactions in chemical space. en. *J. Med. Chem.* **52,** 3103–3107 (May 2009) (cit. on p. 111).

30. McCarroll, M. N., Gendelev, L., Kinser, R., Taylor, J., Bruni, G., Myers-Turnbull, D., Helsell, C., Carbajal, A., Rinaldi, C., Kang, H. J., Gong, J. H., Sello, J. K., Tomita,

S., Peterson, R. T., Keiser, M. J. & Kokel, D. Zebrafish behavioural profiling identifies GABA and serotonin receptor ligands related to sedation and paradoxical excitation. en. *Nat. Commun.* **10,** 4078 (Sept. 2019) (cit. on p. 111).

31. Iwamoto, M., Björklund, T., Lundberg, C., Kirik, D. & Wandless, T. J. A general chemical method to regulate protein stability in the mammalian central nervous system. en. *Chem. Biol.* **17,** 981–988 (Sept. 2010) (cit. on p. 113).

32. Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E. & Tsien, R. Y. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. en. *Nat. Biotechnol.* **22,** 1567–1572 (Dec. 2004) (cit. on p. 113).

33. Xie, C., Soeda, Y., Shinzaki, Y., In, Y., Tomoo, K., Ihara, Y. & Miyasaka, T. Identification of key amino acids responsible for the distinct aggregation properties of microtubule-associated protein 2 and tau. en. *J. Neurochem.* **135,** 19–26 (Oct. 2015) (cit. on p. 113).

34. *Index of /pub/databases/chembl/ChEMBLdb/releases/chembl_26/* `https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_26/`. Accessed: 2021-5-26 (cit. on p. 115).

35. Irwin, J. J., Gaskins, G., Sterling, T., Mysinger, M. M. & Keiser, M. J. Predicted Biological Activity of Purchasable Chemical Space. en. *J. Chem. Inf. Model.* **58,** 148–164 (Jan. 2018) (cit. on p. 115).

36. Benjamini, Y. & Hochberg, Y. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* 1995 (cit. on p. 116).

Appendix A

Supplementary information for Chapter 1

This section presents the Supplementary Information for "Adding Stochastic Negative Examples into Machine Learning Improves Molecular Bioactivity Prediction" (Chapter 1).

A.1 Supporting Methods

A.1.1 SNA + SEAblocklisting

We were concerned over the potential impact of choosing structurally similar ligand analogs as negative examples for SNA. We trained an additional SNA model where predictions from Similarity Ensemble Approach (SEA) for our dataset were blocklisted from SNA selection if SEA predicted likely binders. Effectively, we took these predictions and did not allow the neural network to see these molecule-protein pairs as random transient negatives during minibatching. From the results in Tables S1,S3 and Figures S1, S2, S5, and S6 we found that SEA blocklisting of SNA choices during training improved performance of SNA on Drug Matrix screening benchmarks and Time Split benchmarks, but the bulk magnitude of the

change was relatively small for all but the mean R^2 reported for Drug Matrix with regression DNNs. Classification results were also improved with very small magnitudes when we applied SEA blocklisting to SNA training. We concluded, therefore, that there exists a small risk for negative choice on overall model performance when using SNA. However, as SEA is also a ligand-based method, we suggest further research into more orthogonal methods of blocklisting.

A.1.2 Negatives Upweighted

To perform class balancing by upweighting negative examples, we calculated training loss as defined in Methods then scaled its components by weights based on the positive-to-negative ratio within the minibatch. We calculated a weight for known negative examples within each protein task, which is defined as the maximum value of 1.0 or the number of positive examples for the given target divided by the number of negative examples for the given target. If the weight was less than 1.0 or Nan (in the case where there are no negative examples), we set the weighting scheme to 1.0. Finally, the loss of each interaction is multiplied by the weighting scheme to upweight the impact of existing negatives.

A.1.3 Butina Scaffold Split

To assess whether a strict scaffold split would produce different results from the Time Split benchmark analysis, we created a scaffold split of all ChEMBL training data. As we performed this analysis retroactively after all other benchmarks, this benchmark operates on the same dataset as Time Split, and thus a model trained on this split cannot be evaluated on the Time Split holdout, or vice versa. Using chemfp, we calculated the Tanimoto distance between ECFP4 1024-bit fingerprints[1]. Then, we used Taylor-Butina clustering with a cutoff of 0.4 for Tanimoto[2,3]. Unassigned singletons were assigned to the cluster belonging to their closest match if a match exists over 0.4 Tanimoto similarity. Then we split the data into

a 20% hold out, assigning the largest clusters to Train first similar to Wu et al.[4]. Cross validation was performed over 5 random assignments of scaffolds without replacement and DNN networks were trained as already defined in Methods. Note that in the interests of time we excluded Fold 1 due to edge case behavior we encountered in it, at outlier batch sizes $\leq$2 during training.

A.2 Supporting Figures and Tables

Table A.1: Mean performance metrics and standard deviation across 5-fold cross for all regression models. Models with stochastic negatives used a 1:1 positive-to-negative ratio.

Dataset	Training Type	mean R^2	R^2 std	mean $AUROC_r$	$AUROC_r$ std	mean $AUPRC_r$	$AUPRC_r$ std
STD		0.1926	0.0186	0.6886	0.0094	0.149	0.0077
STD scrambled		0.0154	0.0092	0.5538	0.0099	0.0816	0.0046
SNA		**0.4269**	**0.0272**	**0.7833**	**0.0059**	**0.4405**	**0.0079**
SNA scrambled		0.0021	0.0023	0.4842	0.0134	0.0687	0.003
Negatives Removed	Drugmatrix	0.1973	0.0176	0.612	0.0076	0.1039	0.0025
Negatives Removed scrambled		0.0065	0.0032	0.5315	0.0052	0.0756	0.0014
Negatives Removed +SNA		**0.4257**	**0.0179**	**0.7848**	**0.0053**	**0.4484**	**0.0061**
Negatives Upweighted		0.2177	0.0192	0.7024	0.0062	0.167	0.0081
SNA +SEA blacklist		**0.4411**	**0.0169**	**0.7858**	**0.0051**	**0.4528**	**0.0069**
STD		0.2152	0.0033	**0.7388**	**0.0024**	**0.9434**	**0.0008**
STD scrambled		0.0513	0.0032	0.634	0.0033	0.9057	0.001
SNA		0.1863	0.0012	0.7133	0.0025	0.9401	0.0006
SNA scrambled		0.002	0.0016	0.4664	0.0106	0.854	0.0032
Negatives Removed	Time Split	**0.2352**	**0.0043**	0.7223	0.005	0.9385	0.0009
Negatives Removed scrambled		0.0547	0.0029	0.6253	0.0053	0.9024	0.0011
Negatives Removed +SNA		0.1774	0.0018	0.7091	0.0025	0.9385	0.0007
Negatives Upweighted		0.2179	0.0064	0.7418	0.0036	0.9444	0.0011
SNA +SEA blacklist		0.1878	0.0021	0.715	0.0012	0.9405	0.0004
STD		**0.637**	**0.0041**	**0.9036**	**0.0016**	0.9837	0.0004
STD scrambled		0.0741	0.0026	0.6584	0.0014	0.92	0.0019
SNA		**0.6428**	**0.0058**	**0.9064**	**0.0026**	**0.9848**	**0.0003**
SNA scrambled		0.0009	0.0004	0.47	0.0053	0.8685	0.0012
Negatives Removed	Test	0.6034	0.0014	0.8362	0.0024	0.9704	0.0009
Negatives Removed scrambled		0.082	0.0024	0.6474	0.0011	0.9167	0.0016
Negatives Removed +SNA		0.6026	0.0082	0.8799	0.0028	0.9795	0.0004
Negatives Upweighted		**0.6268**	**0.0057**	0.9018	0.0017	0.9835	0.0003
SNA +SEA blacklist		**0.6462**	**0.0066**	**0.907**	**0.0024**	**0.9849**	**0.0003**
STD		**0.9224**	**0.0095**	**0.9809**	**0.0026**	**0.9972**	**0.0004**
STD scrambled		**0.9212**	**0.0016**	**0.9814**	**0.0002**	**0.9973**	**0.0000**
SNA		0.8971	0.01	0.975	0.0025	0.9962	0.0004
SNA scrambled		0.8618	0.0217	0.9725	0.0047	0.9958	0.0007
Negatives Removed	Train	0.7842	0.0017	0.8223	0.0012	0.9684	0.0002
Negatives Removed scrambled		0.6454	0.0033	0.6518	0.0021	0.9283	0.0007
Negatives Removed +SNA		0.7875	0.0129	0.9127	0.0027	0.9848	0.0005
Negatives Upweighted		0.8712	0.0163	0.9692	0.0041	0.9954	0.0006
SNA +SEA blacklist		0.9064	0.0069	0.9771	0.0015	0.9965	0.0002

Table A.2: Mean performance metrics and standard deviation across 5-fold cross for all classification models. Models with stochastic negatives used a 1:1 positive-to-negative ratio

Dataset	Training Type	mean AUROC	AUROC std	mean AUPRC	AUPRC std
STD (classifier)		0.7202	0.0050	0.1690	0.0024
STD scrambled (classifier)		0.6070	0.0107	0.1075	0.0063
SNA (classifier)		**0.8168**	**0.0047**	**0.4240**	**0.0085**
SNA scrambled (classifier)		0.5645	0.0044	0.0845	0.0029
Negatives Removed (classifier)	Drug Matrix	0.5434	0.0194	0.0794	0.0051
Negatives Removed scrambled (classifier)		0.5275	0.0086	0.0748	0.0033
Negatives Removed +SNA (classifier)		0.8035	0.0034	0.3103	0.0074
Negatives Removed +SNA scrambled (classifier)		0.5440	0.0062	0.0936	0.0015
SNA +SEA blacklist (classifier)		**0.8199**	**0.0034**	**0.4319**	**0.0054**
STD (classifier)		**0.7314**	**0.0044**	**0.9397**	**0.0010**
STD scrambled (classifier)		0.6955	0.0088	0.9259	0.0028
SNA (classifier)		0.7010	0.0016	0.9346	0.0006
SNA scrambled (classifier)		0.6579	0.0029	0.9144	0.0022
Negatives Removed (classifier)	Time Split	0.6332	0.0100	0.9099	0.0037
Negatives Removed scrambled (classifier)		0.6262	0.0056	0.9069	0.0024
Negatives Removed +SNA (classifier)		0.6542	0.0031	0.9187	0.0011
Negatives Removed +SNA scrambled (classifier)		0.5739	0.0055	0.8893	0.0018
SNA +SEA blacklist (classifier)		0.7031	0.0018	0.9354	0.0005
STD (classifier)		**0.9044**	**0.0019**	**0.9827**	**0.0001**
STD scrambled (classifier)		0.7401	0.0038	0.9419	0.0022
SNA (classifier)		0.9010	0.0014	**0.9823**	**0.0003**
SNA scrambled (classifier)		0.7354	0.0035	0.9407	0.0012
Negatives Removed (classifier)	Test	0.6642	0.0075	0.9238	0.0026
Negatives Removed scrambled (classifier)		0.6255	0.0088	0.9111	0.0014
Negatives Removed +SNA (classifier)		0.7091	0.0048	0.9343	0.0020
Negatives Removed +SNA scrambled (classifier)		0.6233	0.0065	0.9110	0.0014
SNA +SEA blacklist (classifier)		0.9004	0.0011	0.9822	0.0003
STD (classifier)		**0.9606**	**0.0033**	**0.9937**	**0.0006**
STD scrambled (classifier)		0.8162	0.0110	0.9665	0.0028
SNA (classifier)		**0.9652**	**0.0023**	**0.9944**	**0.0004**
SNA scrambled (classifier)		0.7473	0.0021	0.9433	0.0016
Negatives Removed (classifier)	Train	0.6655	0.0106	0.9241	0.0026
Negatives Removed scrambled (classifier)		0.6264	0.0052	0.9113	0.0019
Negatives Removed +SNA (classifier)		0.7178	0.0028	0.9359	0.0008
Negatives Removed +SNA scrambled (classifier)		0.6232	0.0047	0.9110	0.0017
SNA +SEA blacklist (classifier)		0.9593	0.0010	0.9933	0.0002

Table A.3: ChEMBL activity relation actions.

ChEMBL activity	Dataset action
'=' or '<'	accept pAC_{50} value as is
'>'	Add np.random 2-3 logs to reported pAC_{50}

Table A.4: Positive and Negative splits for Validation, Train, Time Split, and Drug Matrix

Dataset	Num Positives	Num Negatives	Total	Percent Positive	Percent Negative
Train All	403409	154826	558235	72.3	27.7
Test Fold 0	80861	31341	112202	72.1	27.9
Train Fold 0	322548	123485	446033	72.3	27.7
Test Fold 1	81373	32890	114263	71.2	28.8
Train Fold 1	322036	121936	443972	72.5	27.5
Test Fold 2	79566	29624	109190	72.9	27.1
Train Fold 2	323843	125202	449045	72.1	27.9
Test Fold 3	80312	29874	110186	72.9	27.1
Train Fold 3	323097	124952	448049	72.1	27.9
Test Fold 4	81297	31097	112394	72.3	27.7
Train Fold 4	322112	123729	445841	72.2	27.8
Time Split	83155	33774	116929	71.1	28.9
Drug Matrix (dose response)	2714	330	3044	89.2	10.8
Drug Matrix (primary)	2714	35451	38165	7.1	92.9

Table A.5: Regression performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	$AUROC_r$	$AUROC_r$ std	$AUPRC_r$	$AUPRC_r$ std	R^2	R^2 std
Drug Matrix	Negs Rem. +SNA	0.0000	0.0	100.0	0.6120	0.0076	0.1039	0.0025	0.1973	0.0176
		0.0753	93.0	7.0	0.7404	0.0059	0.3937	0.0043	0.3529	0.0139
		0.1111	90.0	10.0	0.7495	0.0053	0.4061	0.0073	0.3654	0.0207
		0.2500	80.0	20.0	0.7671	0.0060	0.4270	0.0070	0.3865	0.0171
		0.4286	70.0	30.0	0.7708	0.0061	0.4308	0.0085	0.3912	0.0193
		0.6666	**60.0**	**40.0**	0.7757	0.0104	**0.4403**	**0.0104**	**0.4060**	**0.0185**
		0.8182	**55.0**	**45.0**	**0.7864**	**0.0082**	**0.4456**	**0.0087**	**0.4245**	**0.0249**
		1.0000	**50.0**	**50.0**	**0.7848**	**0.0053**	**0.4479**	**0.0062**	**0.4257**	**0.0179**
		1.2222	**45.0**	**55.0**	**0.7937**	**0.0078**	**0.4534**	**0.0099**	**0.4327**	**0.0226**
		1.5000	**40.0**	**60.0**	**0.7969**	**0.0115**	**0.4528**	**0.0106**	**0.4317**	**0.0202**
		2.3333	**30.0**	**70.0**	0.7798	0.0053	0.3695	0.0186	**0.4321**	**0.0139**
		4.0000	20.0	80.0	0.7358	0.0055	0.2457	0.0165	0.3966	0.0163
		9.0000	10.0	90.0	0.6482	0.0070	0.1195	0.0052	0.2816	0.0150
		19.0000	5.0	95.0	0.6124	0.0077	0.1042	0.0054	0.2033	0.0087
	SNA	0.0000	0.0	100.0	0.6886	0.0094	0.1490	0.0077	0.1926	0.0186
		0.0753	93.0	7.0	0.7474	0.0103	0.4032	0.0070	0.3641	0.0215
		0.1111	90.0	10.0	0.7540	0.0098	0.4166	0.0087	0.3800	0.0202
		0.2500	80.0	20.0	0.7697	0.0070	0.4380	0.0048	0.4028	0.0202
		0.4286	**70.0**	**30.0**	0.7721	0.0078	0.4367	0.0057	**0.4092**	**0.0203**
		0.6666	**60.0**	**40.0**	0.7799	0.0034	**0.4433**	**0.0018**	**0.4187**	**0.0166**
		0.8182	**55.0**	**45.0**	**0.7857**	**0.0064**	**0.4453**	**0.0040**	**0.4239**	**0.0166**
		1.0000	**50.0**	**50.0**	**0.7841**	**0.0061**	**0.4428**	**0.0054**	**0.4282**	**0.0215**
		1.2222	**45.0**	**55.0**	**0.7901**	**0.0032**	**0.4502**	**0.0024**	**0.4329**	**0.0151**
		1.5000	**40.0**	**60.0**	**0.7908**	**0.0032**	**0.4490**	**0.0062**	**0.4341**	**0.0159**
		2.3333	**30.0**	**70.0**	0.7803	0.0066	0.3702	0.0194	**0.4383**	**0.0194**
		4.0000	20.0	80.0	0.7434	0.0066	0.2504	0.0159	0.3987	0.0093
		9.0000	10.0	90.0	0.7012	0.0064	0.1661	0.0065	0.3184	0.0152
		19.0000	5.0	95.0	0.6831	0.0063	0.1477	0.0042	0.2172	0.0229
		0.0000	0.0	100.0	0.8362	0.0024	0.9704	0.0009	0.6034	0.0014
		0.0753	93.0	7.0	0.8511	0.0021	0.9741	0.0003	0.4798	0.0053
		0.1111	90.0	10.0	0.8600	0.0023	0.9760	0.0003	0.5137	0.0066
		0.2500	80.0	20.0	0.8730	0.0025	0.9784	0.0004	0.5631	0.0072
		0.4286	**70.0**	**30.0**	**0.8782**	**0.0024**	**0.9793**	**0.0002**	0.5852	0.0055

Table A.5: Regression performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	$AUROC_r$	$AUROC_r$ std	$AUPRC_r$	$AUPRC_r$ std	R^2	R^2 std
Test	Negs Rem. +SNA	**0.6666**	**60.0**	**40.0**	**0.8790**	**0.0027**	**0.9795**	**0.0004**	0.5910	0.0085
		0.8182	**55.0**	**45.0**	**0.8779**	**0.0027**	**0.9792**	**0.0004**	0.5966	0.0051
		1.0000	**50.0**	**50.0**	**0.8799**	**0.0028**	**0.9795**	**0.0004**	0.6026	0.0082
		1.2222	**45.0**	**55.0**	**0.8801**	**0.0033**	**0.9796**	**0.0005**	0.6036	0.0079
		1.5000	**40.0**	**60.0**	**0.8793**	**0.0018**	**0.9794**	**0.0004**	0.6040	0.0053
		2.3333	**30.0**	**70.0**	0.8745	**0.0035**	**0.9783**	**0.0010**	**0.6203**	**0.0032**
		4.0000	**20.0**	**80.0**	0.8649	0.0024	0.9764	0.0007	**0.6205**	**0.0019**
		9.0000	10.0	90.0	0.8465	0.0023	0.9727	0.0008	0.6084	0.0021
		19.0000	5.0	95.0	0.8358	0.0028	0.9705	0.0008	0.6017	0.0024
	SNA	0.0000	0.0	100.0	0.9036	0.0016	0.9837	0.0004	0.6370	0.0041
		0.0753	93.0	7.0	0.8674	0.0020	0.9778	0.0002	0.5122	0.0044
		0.1111	90.0	10.0	0.8799	0.0024	0.9802	0.0002	0.5521	0.0063
		0.2500	80.0	20.0	0.8962	0.0022	0.9831	0.0002	0.6074	0.0068
		0.4286	70.0	30.0	0.9022	0.0028	0.9841	0.0003	0.6277	0.0068
		0.6666	60.0	40.0	0.9035	0.0027	0.9843	0.0002	0.6328	0.0066
		0.8182	**55.0**	**45.0**	**0.9055**	**0.0024**	**0.9846**	**0.0003**	0.6408	0.0049
		1.0000	**50.0**	**50.0**	**0.9061**	**0.0025**	**0.9847**	**0.0003**	**0.6419**	**0.0054**
		1.2222	**45.0**	**55.0**	**0.9064**	**0.0023**	**0.9848**	**0.0003**	**0.6432**	**0.0066**
		1.5000	**40.0**	**60.0**	**0.9070**	**0.0027**	**0.9849**	**0.0002**	**0.6451**	**0.0076**
		2.3333	**30.0**	**70.0**	**0.9085**	**0.0012**	**0.9849**	**0.0003**	**0.6520**	**0.0034**
		4.0000	**20.0**	**80.0**	**0.9076**	**0.0019**	**0.9846**	**0.0004**	**0.6516**	**0.0039**
		9.0000	**10.0**	**90.0**	**0.9069**	**0.0027**	**0.9844**	**0.0004**	**0.6509**	**0.0051**
		19.0000	**5.0**	**95.0**	**0.9066**	**0.0024**	0.9842	0.0005	**0.6501**	**0.0043**
		0.0000	**0.0**	**100.0**	**0.7223**	**0.0050**	**0.9385**	**0.0009**	**0.2352**	**0.0043**
		0.0753	93.0	7.0	0.7025	0.0016	0.9360	0.0004	0.1608	0.0013
		0.1111	90.0	10.0	0.7011	0.0008	0.9359	0.0002	0.1616	0.0013
		0.2500	80.0	20.0	0.7032	0.0019	0.9367	0.0004	0.1677	0.0011
		0.4286	70.0	30.0	0.7028	0.0011	0.9367	0.0002	0.1691	0.0017

Table A.5: Regression performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	$AUROC_r$	$AUROC_r$ std	$AUPRC_r$	$AUPRC_r$ std	R^2	R^2 std
	Negs Rem. +SNA	0.6666	60.0	40.0	0.7049	0.0014	0.9374	0.0004	0.1720	0.0016
		0.8182	55.0	45.0	0.7070	0.0023	0.9379	0.0006	0.1764	0.0018
		1.0000	50.0	50.0	0.7091	0.0025	0.9385	0.0007	0.1774	0.0018
		1.2222	45.0	55.0	0.7098	0.0026	0.9387	0.0006	0.1793	0.0035
		1.5000	40.0	60.0	0.7111	0.0019	0.9390	0.0004	0.1792	0.0012
		2.3333	30.0	70.0	0.6973	0.0012	0.9356	0.0004	0.1820	0.0027
		4.0000	20.0	80.0	0.7005	0.0025	0.9347	0.0007	0.1899	0.0052
		9.0000	10.0	90.0	0.7166	0.0021	0.9380	0.0007	0.2163	0.0040
Time		**19.0000**	**5.0**	**95.0**	**0.7277**	**0.0056**	**0.9398**	**0.0012**	**0.2437**	**0.0061**
Split		**0.0000**	**0.0**	**100.0**	**0.7388**	**0.0024**	**0.9434**	**0.0008**	**0.2152**	**0.0033**
	SNA	0.0753	93.0	7.0	0.7015	0.0012	0.9369	0.0004	0.1682	0.0011
		0.1111	90.0	10.0	0.7037	0.0015	0.9377	0.0004	0.1731	0.0015
		0.2500	80.0	20.0	0.7060	0.0020	0.9383	0.0006	0.1785	0.0017
		0.4286	70.0	30.0	0.7090	0.0016	0.9391	0.0005	0.1825	0.0028
		0.6666	60.0	40.0	0.7104	0.0033	0.9392	0.0007	0.1835	0.0018
		0.8182	55.0	45.0	0.7113	0.0023	0.9395	0.0006	0.1846	0.0024
		1.0000	50.0	50.0	0.7109	0.0023	0.9395	0.0006	0.1831	0.0028
		1.2222	45.0	55.0	0.7122	0.0029	0.9398	0.0007	0.1838	0.0022
		1.5000	40.0	60.0	0.7129	0.0023	0.9400	0.0006	0.1843	0.0023
		2.3333	30.0	70.0	0.7131	0.0033	0.9398	0.0007	0.1911	0.0024
		4.0000	20.0	80.0	0.7210	0.0022	0.9406	0.0005	0.1997	0.0036
		9.0000	**10.0**	**90.0**	**0.7329**	**0.0018**	**0.9426**	**0.0004**	**0.2154**	**0.0021**
		19.0000	**5.0**	**95.0**	**0.7380**	**0.0041**	**0.9431**	**0.0011**	**0.2213**	**0.0050**
		0.0000	0.0	100.0	0.8223	0.0012	0.9684	0.0002	0.7842	0.0017
		0.0753	93.0	7.0	0.8775	0.0014	0.9787	0.0002	0.5620	0.0036
		0.1111	90.0	10.0	0.8905	0.0015	0.9811	0.0003	0.6236	0.0043
		0.2500	**80.0**	**20.0**	**0.9090**	**0.0017**	0.9843	0.0003	0.7247	0.0037
		0.4286	**70.0**	**30.0**	**0.9154**	**0.0017**	**0.9853**	**0.0003**	0.7744	0.0097

Table A.5: Regression performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	$AUROC_r$	$AUROC_r$ std	$AUPRC_r$	$AUPRC_r$ std	R^2	R^2 std
Train	Negs Rem +SNA	**0.6666**	**60.0**	**40.0**	**0.9158**	**0.0028**	**0.9854**	**0.0004**	**0.7835**	**0.0118**
		0.8182	**55.0**	**45.0**	**0.9095**	**0.0020**	0.9842	0.0004	0.7726	0.0108
		1.0000	**50.0**	**50.0**	**0.9127**	**0.0027**	**0.9848**	**0.0005**	**0.7875**	**0.0129**
		1.2222	**45.0**	**55.0**	**0.9123**	**0.0038**	**0.9847**	**0.0006**	**0.7887**	**0.0166**
		1.5000	**40.0**	**60.0**	**0.9104**	**0.0017**	0.9843	0.0003	0.7840	0.0088
		2.3333	**30.0**	**70.0**	0.8950	0.0036	0.9817	0.0005	**0.8049**	**0.0133**
		4.0000	**20.0**	**80.0**	0.8756	0.0016	0.9783	0.0003	**0.8074**	**0.0062**
		9.0000	**10.0**	**90.0**	0.8450	0.0007	0.9727	0.0002	**0.7918**	**0.0085**
		19.0000	**5.0**	**95.0**	0.8240	0.0021	0.9688	0.0005	**0.7840**	**0.0039**
	SNA	**0.0000**	**0.0**	**100.0**	**0.9809**	**0.0026**	0.9972	0.0004	**0.9224**	**0.0095**
		0.0753	93.0	7.0	0.8966	0.0006	0.9830	0.0001	0.5984	0.0011
		0.1111	90.0	10.0	0.9156	0.0007	0.9864	0.0002	0.6670	0.0021
		0.2500	80.0	20.0	0.9462	0.0009	0.9916	0.0002	0.7859	0.0040
		0.4286	70.0	30.0	0.9623	0.0010	0.9942	0.0001	0.8487	0.0028
		0.6666	60.0	40.0	0.9673	0.0009	0.9950	0.0001	0.8687	0.0025
		0.8182	55.0	45.0	0.9733	0.0012	0.9959	0.0002	0.8912	0.0048
		1.0000	50.0	50.0	0.9753	0.0017	0.9962	0.0003	0.8984	0.0061
		1.2222	45.0	55.0	0.9746	0.0018	0.9961	0.0003	0.8961	0.0074
		1.5000	40.0	60.0	0.9762	0.0016	0.9964	0.0002	0.9022	0.0075
		2.3333	**30.0**	**70.0**	**0.9832**	**0.0017**	0.9975	0.0003	**0.9287**	**0.0062**
		4.0000	**20.0**	**80.0**	**0.9861**	**0.0007**	**0.9979**	**0.0001**	**0.9399**	**0.0032**
		9.0000	**10.0**	**90.0**	**0.9844**	**0.0026**	**0.9977**	**0.0004**	**0.9339**	**0.0097**
		19.0000	**5.0**	**95.0**	**0.9830**	**0.0042**	**0.9974**	**0.0006**	**0.9296**	**0.0165**

Table A.6: Classification performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	AUROC	AUROC std	AUPRC	AUPRC std
Drug Matrix	Negs Rem. +SNA	0.0000	0.0	100.0	0.5434	0.0194	0.0794	0.0051
		0.0753	93.0	7.0	0.7391	0.0022	0.2235	0.0012
		0.1111	90.0	10.0	0.7515	0.0020	0.2570	0.0016
		0.2500	80.0	20.0	0.7787	0.0089	0.2667	0.0158
		0.4286	70.0	30.0	0.7934	0.0026	0.2940	0.0066
		0.6666	**60.0**	**40.0**	0.8003	0.0025	**0.3035**	**0.0038**
		0.8182	**55.0**	**45.0**	**0.8014**	**0.0034**	**0.3138**	**0.0048**
		1.0000	**50.0**	**50.0**	**0.8035**	**0.0034**	**0.3103**	**0.0074**
		1.2222	**45.0**	**55.0**	**0.8064**	**0.0037**	**0.3116**	**0.0065**
		1.5000	**40.0**	**60.0**	**0.8086**	**0.0035**	**0.3060**	**0.0091**
		2.3333	**30.0**	**70.0**	0.7690	0.0040	0.2832	0.0104
		4.0000	20.0	80.0	0.6846	0.0041	0.1558	0.0120
		9.0000	10.0	90.0	0.5331	0.0188	0.0834	0.0060
		19.0000	5.0	95.0	0.5466	0.0059	0.0777	0.0012
	SNA	0.0000	0.0	100.0	0.7202	0.0050	0.1690	0.0024
		0.0753	93.0	7.0	0.7870	0.0034	0.3258	0.0038
		0.1111	90.0	10.0	0.7952	0.0048	0.3485	0.0051
		0.2500	80.0	20.0	0.8095	0.0028	0.4038	0.0059
		0.4286	**70.0**	**30.0**	0.8113	0.0038	0.4159	0.0050
		0.6666	**60.0**	**40.0**	0.8167	0.0017	**0.4235**	**0.0041**
		0.8182	**55.0**	**45.0**	**0.8119**	**0.0040**	**0.4176**	**0.0085**
		1.0000	**50.0**	**50.0**	**0.8168**	**0.0047**	**0.4240**	**0.0085**
		1.2222	**45.0**	**55.0**	**0.8216**	**0.0042**	**0.4280**	**0.0064**
		1.5000	**40.0**	**60.0**	**0.8228**	**0.0038**	**0.4248**	**0.0041**
		2.3333	**30.0**	**70.0**	0.8051	0.0025	0.3924	0.0074
		4.0000	20.0	80.0	0.7789	0.0033	0.3115	0.0105
		9.0000	10.0	90.0	0.7470	0.0028	0.2151	0.0104
		19.0000	5.0	95.0	0.7136	0.0019	0.1637	0.0041
		0.0000	0.0	100.0	0.6642	0.0075	0.9238	0.0026
		0.0753	93.0	7.0	0.6342	0.0021	0.9226	0.0020
		0.1111	90.0	10.0	0.6713	0.0017	0.9273	0.0016
		0.2500	80.0	20.0	0.6930	0.0044	0.9306	0.0019
		0.4286	**70.0**	**30.0**	**0.7018**	**0.0033**	**0.9324**	**0.0010**

Table A.6: Classification performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	AUROC	AUROC std	AUPRC	AUPRC std
Test	Negs Rem. +SNA	**0.6666**	**60.0**	**40.0**	**0.7057**	**0.0033**	**0.9332**	**0.0022**
		0.8182	**55.0**	**45.0**	**0.7085**	**0.0028**	**0.9340**	**0.0014**
		1.0000	**50.0**	**50.0**	**0.7091**	**0.0048**	**0.9343**	**0.0020**
		1.2222	**45.0**	**55.0**	**0.7097**	**0.0034**	**0.9344**	**0.0015**
		1.5000	**40.0**	**60.0**	**0.7073**	**0.0040**	**0.9338**	**0.0024**
		2.3333	**30.0**	**70.0**	0.6774	**0.0045**	**0.9296**	**0.0016**
		4.0000	**20.0**	**80.0**	0.6177	0.0059	0.9122	0.0028
		9.0000	10.0	90.0	0.5692	0.0154	0.8976	0.0034
		19.0000	5.0	95.0	0.6538	0.0073	0.9201	0.0016
	SNA	**0.0000**	**0.0**	**100.0**	**0.9044**	**0.0019**	0.9827	0.0001
		0.0753	93.0	7.0	0.8074	0.0017	0.9626	0.0005
		0.1111	90.0	10.0	0.8297	0.0066	0.9677	0.0014
		0.2500	80.0	20.0	0.8706	0.0019	0.9763	0.0004
		0.4286	70.0	30.0	0.8881	0.0019	0.9799	0.0003
		0.6666	60.0	40.0	0.8926	0.0018	0.9807	0.0002
		0.8182	**55.0**	**45.0**	**0.8999**	**0.0015**	**0.9821**	**0.0003**
		1.0000	**50.0**	**50.0**	**0.9010**	**0.0014**	**0.9823**	**0.0003**
		1.2222	**45.0**	**55.0**	**0.8993**	**0.0072**	**0.9820**	**0.0014**
		1.5000	**40.0**	**60.0**	**0.9030**	**0.0013**	**0.9827**	**0.0004**
		2.3333	**30.0**	**70.0**	**0.9063**	**0.0013**	**0.9833**	**0.0004**
		4.0000	**20.0**	**80.0**	**0.9057**	**0.0024**	**0.9831**	**0.0004**
		9.0000	**10.0**	**90.0**	**0.9052**	**0.0023**	**0.9829**	**0.0003**
		19.0000	**5.0**	**95.0**	**0.9031**	**0.0032**	0.9825	0.0003
		0.0000	**0.0**	**100.0**	**0.6332**	**0.0100**	**0.9099**	**0.0037**
		0.0753	93.0	7.0	0.5904	0.0007	0.9036	0.0006
		0.1111	90.0	10.0	0.6339	0.0009	0.9134	0.0003
		0.2500	80.0	20.0	0.6454	0.0037	0.9168	0.0015
		0.4286	**70.0**	**30.0**	0.6534	0.0018	**0.9192**	**0.0003**

Table A.6: Classification performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	AUROC	AUROC std	AUPRC	AUPRC std
Time Split	Negs Rem. +SNA	**0.6666**	**60.0**	**40.0**	**0.6554**	**0.0019**	**0.9193**	**0.0007**
		0.8182	**55.0**	**45.0**	**0.6569**	**0.0013**	**0.9197**	**0.0003**
		1.0000	**50.0**	**50.0**	**0.6542**	**0.0031**	**0.9187**	**0.0011**
		1.2222	**45.0**	**55.0**	**0.6574**	**0.0013**	**0.9194**	**0.0001**
		1.5000	**40.0**	**60.0**	**0.6546**	**0.0028**	**0.9189**	**0.0006**
		2.3333	30.0	70.0	0.6247	0.0015	0.9123	0.0008
		4.0000	20.0	80.0	0.5893	0.0054	0.8951	0.0017
		9.0000	10.0	90.0	0.5745	0.0064	0.8916	0.0041
		19.0000	**5.0**	**95.0**	**0.6259**	**0.0086**	**0.9058**	**0.0035**
	SNA	**0.0000**	**0.0**	**100.0**	**0.7314**	**0.0044**	**0.9397**	**0.0010**
		0.0753	93.0	7.0	0.6791	0.0012	0.9281	0.0003
		0.1111	90.0	10.0	0.6829	0.0010	0.9294	0.0003
		0.2500	80.0	20.0	0.6890	0.0006	0.9320	0.0003
		0.4286	70.0	30.0	0.6954	0.0007	0.9336	0.0002
		0.6666	60.0	40.0	0.6950	0.0011	0.9333	0.0005
		0.8182	55.0	45.0	0.6982	0.0026	0.9338	0.0007
		1.0000	50.0	50.0	0.7010	0.0016	0.9346	0.0006
		1.2222	45.0	55.0	0.7029	0.0040	0.9350	0.0012
		1.5000	40.0	60.0	0.7012	0.0023	0.9343	0.0008
		2.3333	30.0	70.0	0.7134	0.0037	0.9374	0.0012
		4.0000	20.0	80.0	0.7226	0.0026	0.9387	0.0009
		9.0000	**10.0**	**90.0**	**0.7302**	**0.0031**	**0.9398**	**0.0007**
		19.0000	**5.0**	**95.0**	**0.7321**	**0.0022**	**0.9401**	**0.0006**
		0.0000	0.0	100.0	0.6655	0.0106	0.9241	0.0026
		0.0753	93.0	7.0	0.6363	0.0019	0.9232	0.0008
		0.1111	90.0	10.0	0.6750	0.0009	0.9281	0.0007
		0.2500	**80.0**	**20.0**	**0.7001**	**0.0022**	0.9321	0.0009
		0.4286	**70.0**	**30.0**	**0.7087**	**0.0040**	**0.9337**	**0.0013**

Table A.6: Classification performance for SNA models across multiple positive to negative ratios.

Dataset	Model	Pos:Neg Ratio	Target Min %Neg	Target %Pos	AUROC	AUROC std	AUPRC	AUPRC std
Train	Negs Rem +SNA	**0.6666**	**60.0**	**40.0**	**0.7138**	**0.0027**	**0.9349**	**0.0007**
		0.8182	**55.0**	**45.0**	**0.7169**	**0.0005**	**0.9356**	**0.0005**
		1.0000	**50.0**	**50.0**	**0.7178**	**0.0028**	**0.9359**	**0.0008**
		1.2222	**45.0**	**55.0**	**0.7184**	**0.0017**	**0.9361**	**0.0008**
		1.5000	**40.0**	**60.0**	**0.7153**	**0.0039**	**0.9352**	**0.0009**
		2.3333	**30.0**	**70.0**	0.6831	0.0039	0.9308	0.0014
		4.0000	**20.0**	**80.0**	0.6204	0.0043	0.9126	0.0015
		9.0000	**10.0**	**90.0**	0.5712	0.0151	0.8981	0.0037
		19.0000	**5.0**	**95.0**	0.6552	0.0086	0.9204	0.0026
	SNA	**0.0000**	**0.0**	**100.0**	**0.9606**	**0.0033**	**0.9937**	**0.0006**
		0.0753	93.0	7.0	0.8277	0.0009	0.9669	0.0002
		0.1111	90.0	10.0	0.8547	0.0075	0.9728	0.0017
		0.2500	80.0	20.0	0.9101	0.0011	0.9841	0.0002
		0.4286	70.0	30.0	0.9387	0.0009	0.9896	0.0002
		0.6666	**60.0**	**40.0**	**0.9471**	**0.0012**	**0.9912**	**0.0002**
		0.8182	**55.0**	**45.0**	**0.9640**	**0.0009**	**0.9942**	**0.0002**
		1.0000	**50.0**	**50.0**	**0.9652**	**0.0023**	**0.9944**	**0.0004**
		1.2222	**45.0**	**55.0**	**0.9586**	**0.0176**	**0.9931**	**0.0033**
		1.5000	**40.0**	**60.0**	**0.9698**	**0.0020**	**0.9952**	**0.0003**
		2.3333	**30.0**	**70.0**	**0.9674**	**0.0032**	**0.9948**	**0.0005**
		4.0000	**20.0**	**80.0**	**0.9648**	**0.0027**	**0.9944**	**0.0005**
		9.0000	**10.0**	**90.0**	**0.9619**	**0.0016**	**0.9939**	**0.0003**
		19.0000	**5.0**	**95.0**	**0.9596**	**0.0054**	**0.9935**	**0.0010**

Table A.7: Regression performance for Butina Split SNA models.

Model	Dataset	mean R^2	R^2 std	mean $AUROC_r$	$AUROC_r$ std	mean $AUPRC_r$	$AUPRC_r$ std
STD	Drug Matrix	0.1547	0.0026	0.6683	0.005	0.1342	0.0029
STD scrambled		0.0112	0.009	0.5196	0.0061	0.0735	0.0012
SNA		**0.3939**	**0.0521**	**0.7863**	**0.0075**	**0.4273**	**0.0109**
SNA scrambled		0.0022	0.0015	0.5027	0.0093	0.0737	0.0017
Negatives Removed		0.1794	0.0226	0.6249	0.0081	0.1078	0.0043
Negatives Removed scrambled		0.0083	0.0094	0.5069	0.0057	0.0707	0.0017
Negatives Removed +SNA		**0.3832**	**0.0388**	**0.7956**	**0.002**	**0.4224**	**0.0182**
Negatives Removed + SNA scrambled		0.0043	0.0049	0.4712	0.0386	0.0717	0.0102
STD	Butina Scaffold Split	**0.3692**	**0.0027**	**0.8092**	**0.0026**	**0.9435**	**0.0007**
STD scrambled		0.0687	0.0027	0.6428	0.0027	0.869	0.0004
SNA		0.3201	0.0021	0.7936	0.0023	**0.9434**	**0.0006**
SNA scrambled		0.0009	0.0005	0.4786	0.0032	0.8074	0.0023
Negatives Removed		0.3411	0.0065	0.7618	0.0018	0.9263	0.0008
Negatives Removed scrambled		0.0841	0.0034	0.652	0.0023	0.871	0.0021
Negatives Removed +SNA		0.2594	0.0201	0.7608	0.0094	0.9305	0.0042
Negatives Removed + SNA scrambled		0.0013	0.0016	0.5101	0.0181	0.8234	0.009
STD	Butina Split Validation	**0.3229**	**0.0111**	**0.8147**	**0.006**	**0.9682**	**0.0029**
STD scrambled		0.0543	0.0072	0.646	0.0108	0.9241	0.0059
SNA		0.2747	0.0169	0.7946	0.0087	**0.967**	**0.0029**
SNA scrambled		0.0004	0.0003	0.4776	0.0025	0.8831	0.0044
Negatives Removed		0.2578	0.022	0.782	0.0072	0.9613	0.0032
Negatives Removed scrambled		0.0531	0.0105	0.656	0.0058	0.9251	0.0037
NEG_RM_SNA		0.219	0.0136	0.7619	0.0116	0.9599	0.0034
NEG_RM_SNA_scrambled		0.001	0.0007	0.5095	0.0168	0.8942	0.0078
STD	Train	0.9411	0.0082	0.9857	0.0019	0.9981	0.0003
STD scrambled		**0.9615**	**0.005**	**0.9909**	**0.0014**	**0.9988**	**0.0002**
SNA		0.8515	0.0276	0.9638	0.0063	0.9949	0.001
SNA scrambled		0.7372	0.0123	0.9453	0.0038	0.9922	0.0007
Negatives Removed		0.7661	0.0113	0.8349	0.0015	0.9731	0.0005
Negatives Removed scrambled		0.5441	0.0091	0.6965	0.0039	0.9406	0.0014
Negatives Removed +SNA		0.6906	0.1007	0.8913	0.0226	0.9824	0.0042
Negatives Removed + SNA scrambled		0.4269	0.0216	0.8472	0.0064	0.9734	0.0016

Table A.8: Classification performance for Butina Split SNA models.

Model	Dataset	mean AUROC	AUROC std	mean AUPRC	AUPRC std
STD (classifier)	Drug Matrix	0.6976	0.0035	0.1610	0.0046
STD scrambled (classifier)		0.5673	0.0119	0.0819	0.0038
SNA (classifier)		**0.8012**	**0.0139**	**0.3660**	**0.0241**
SNA scrambled (classifier)		0.5547	0.0104	0.0796	0.0043
Negatives Removed (classifier)		0.5763	0.0088	0.0887	0.0024
Negatives Removed scrambled (classifier)		0.5514	0.0100	0.0821	0.0027
Negatives Removed +SNA (classifier)		**0.7960**	**0.0060**	0.3115	0.0041
Negatives Removed +SNA scrambled (classifier)		0.5414	0.0116	0.0823	0.0043
STD (classifier)	Butina Scaffold Split	**0.7955**	**0.0010**	**0.9376**	**0.0004**
STD scrambled (classifier)		0.7117	0.0023	0.9012	0.0019
SNA (classifier)		0.7362	0.0073	0.9186	0.0030
Negatives Removed (classifier)		0.6674	0.0028	0.8776	0.0024
SNA scrambled (classifier)		0.6558	0.0016	0.8772	0.0015
Negatives Removed scrambled (classifier)		0.6574	0.0032	0.8765	0.0020
Negatives Removed +SNA (classifier)		0.6522	0.0009	0.8812	0.0011
Negatives Removed +SNA scrambled (classifier)		0.5714	0.0078	0.8427	0.0046
STD (classifier)	Butina Split Validation	**0.7987**	**0.0169**	**0.9639**	**0.0048**
STD scrambled (classifier)		0.7156	0.0199	0.9424	0.0073
SNA (classifier)		0.7463	0.0125	0.9534	0.0045
SNA scrambled (classifier)		0.6998	0.0188	0.9346	0.0095
Negatives Removed (classifier)		0.6489	0.0082	0.9267	0.0056
Negatives Removed scrambled (classifier)		0.6044	0.0129	0.9135	0.0064
Negatives Removed +SNA (classifier)		0.6537	0.0225	0.9276	0.0088
Negatives Removed +SNA scrambled (classifier)		0.5802	0.0083	0.9080	0.0061
STD (classifier)	Train	**0.9093**	**0.0047**	**0.9855**	**0.0010**
STD scrambled (classifier)		0.7747	0.0036	0.9576	0.0013
SNA (classifier)		0.8601	0.0174	0.9758	0.0035
SNA scrambled (classifier)		0.7588	0.0043	0.9527	0.0013
Negatives Removed (classifier)		0.6781	0.0038	0.9331	0.0018
Negatives Removed scrambled (classifier)		0.6333	0.0041	0.9203	0.0023
Negatives Removed +SNA (classifier)		0.7197	0.0066	0.9420	0.0022
Negatives Removed +SNA scrambled (classifier)		0.6090	0.0052	0.9159	0.0024

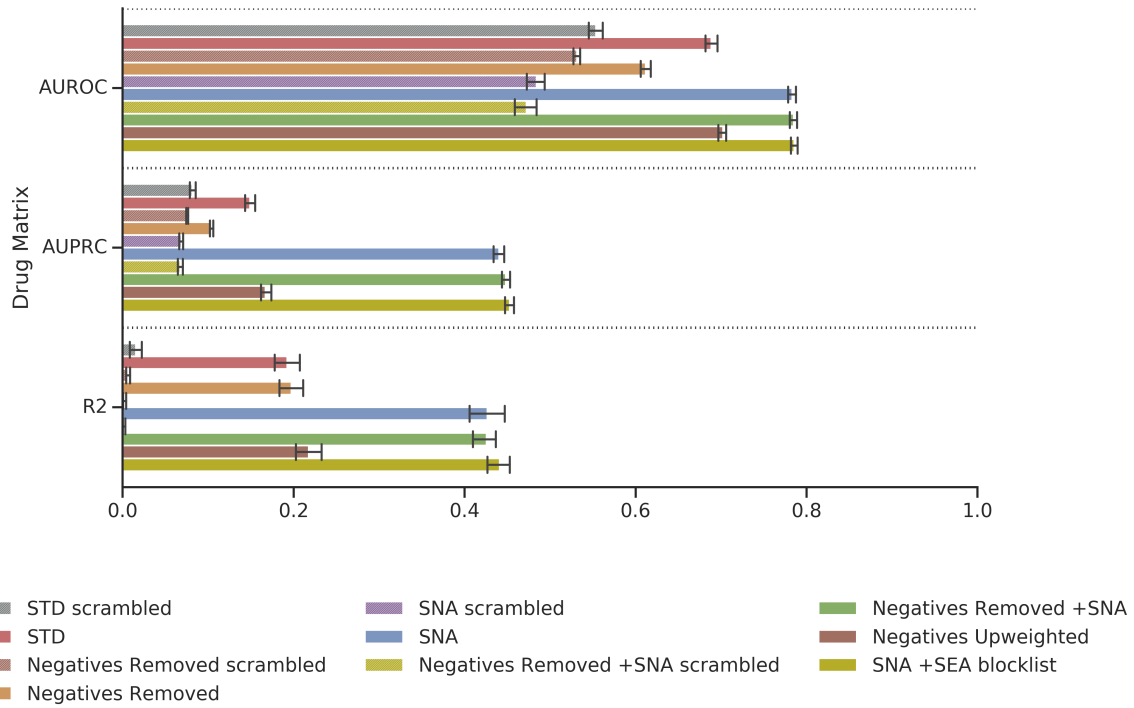

Figure A.1: Drug Matrix performance for all regression models. Note: R^2 values for SNA scrambled and Negatives Removed +SNA are close to 0.0

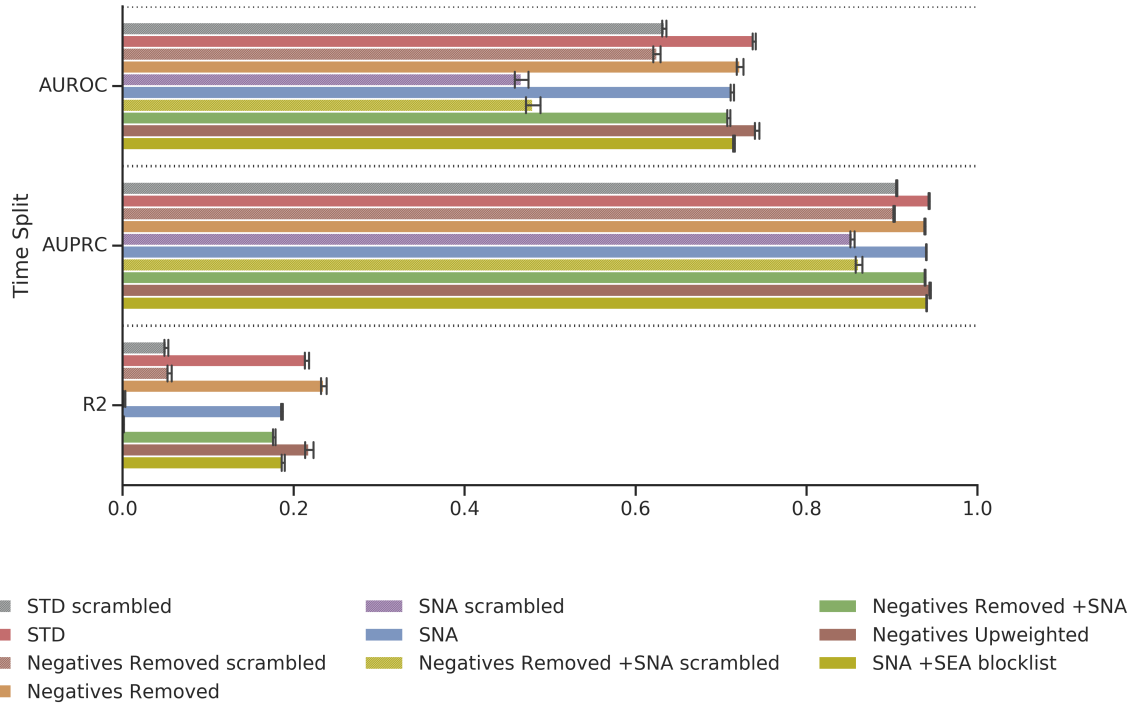

Figure A.2: Time Split performance for all regression models. Note: R^2 values for SNA scrambled and Negatives Removed +SNA are close to 0.0

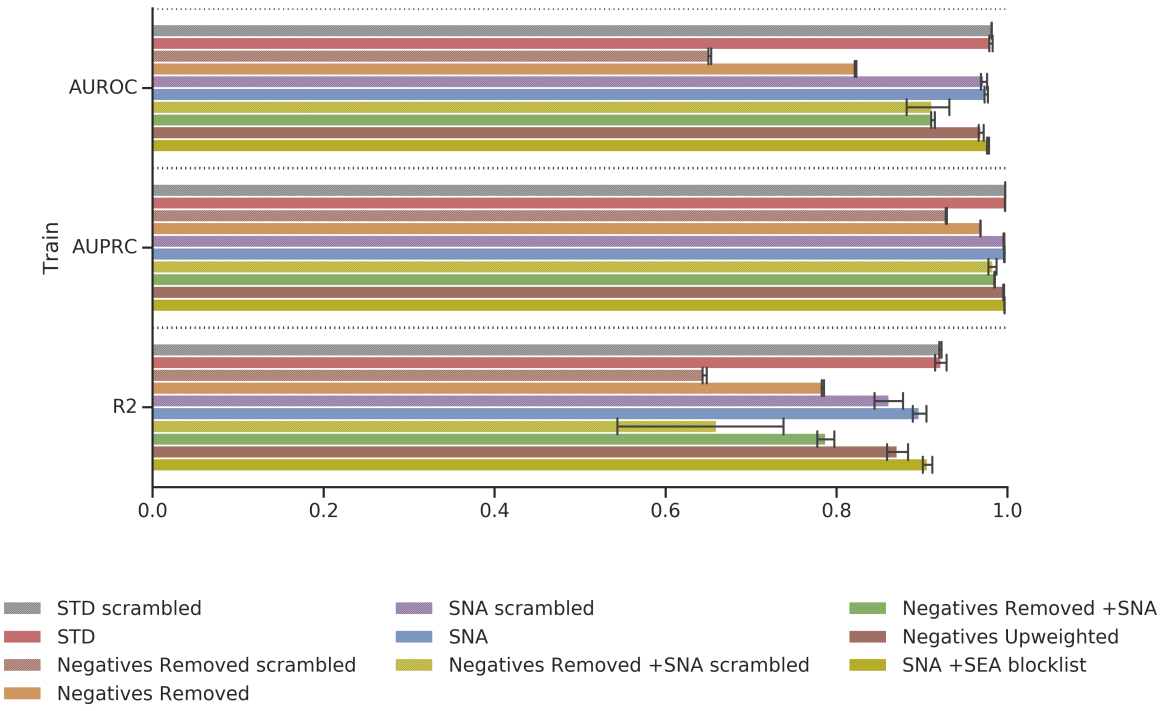

Figure A.3: Train performance for all regression models

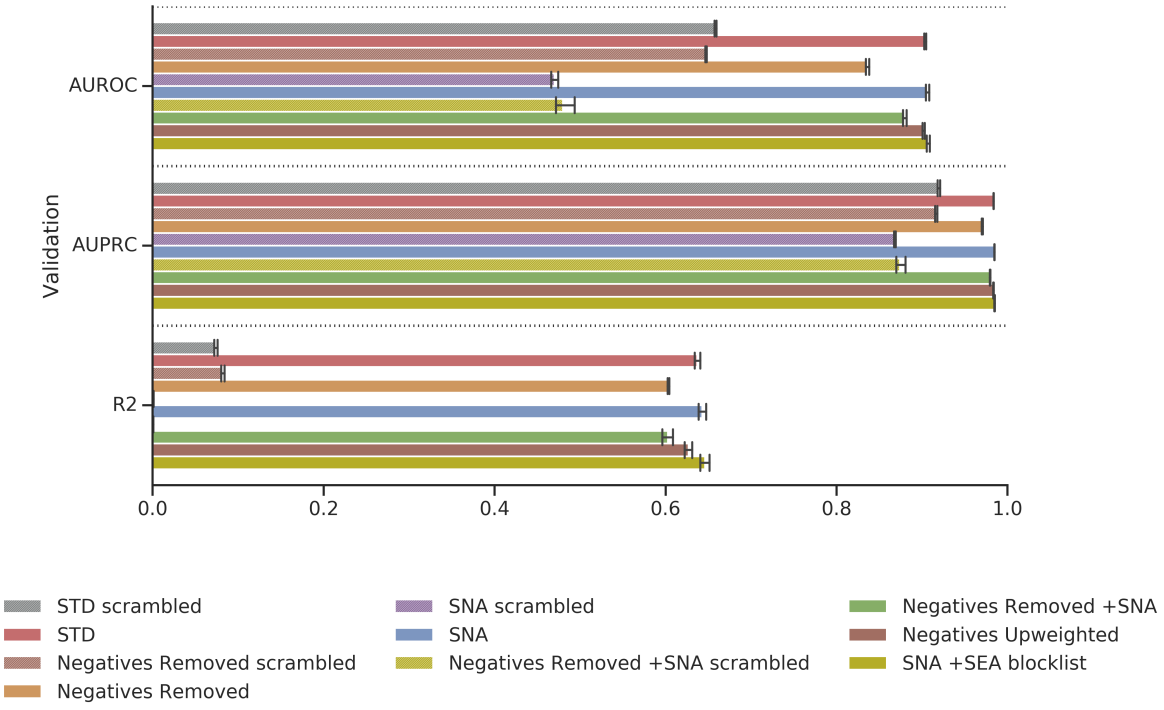

Figure A.4: Validation performance for all regression

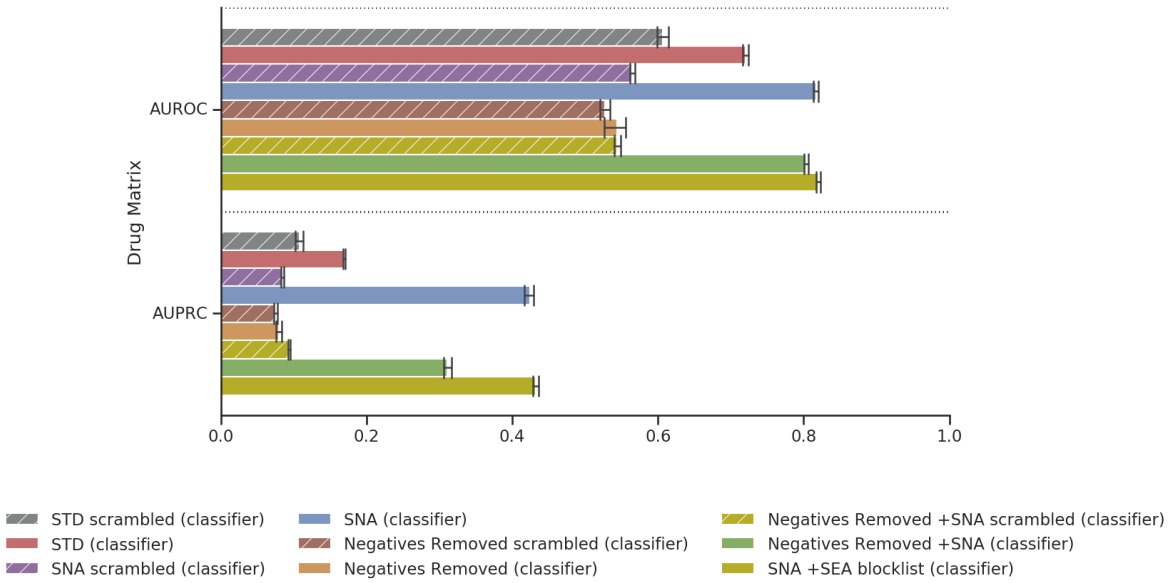

Figure A.5: Drug Matrix performance for all classification models

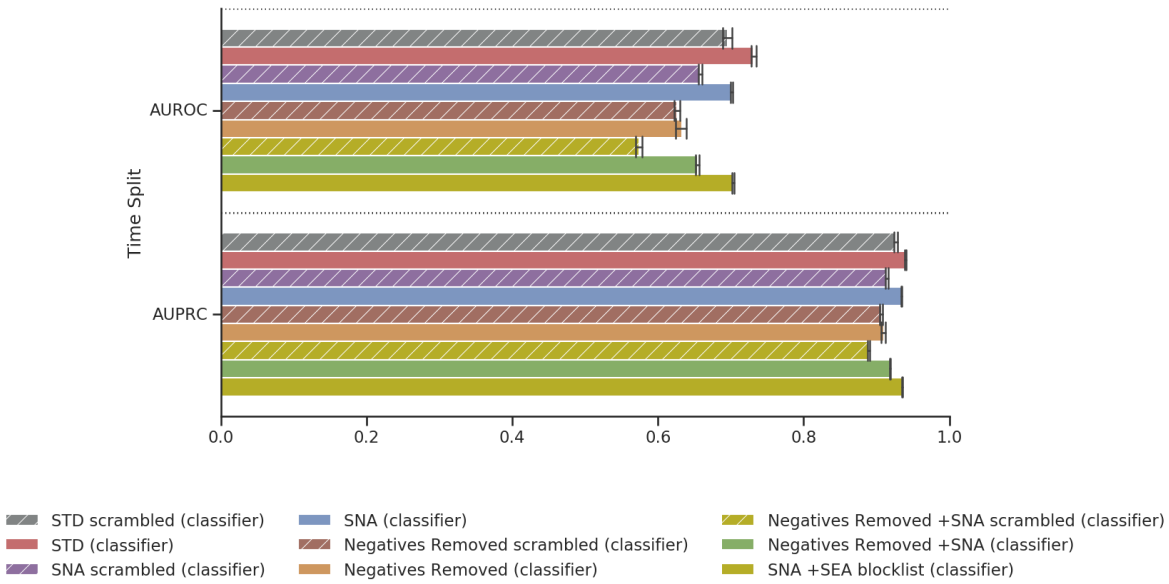

Figure A.6: Time Split performance for all classification models.

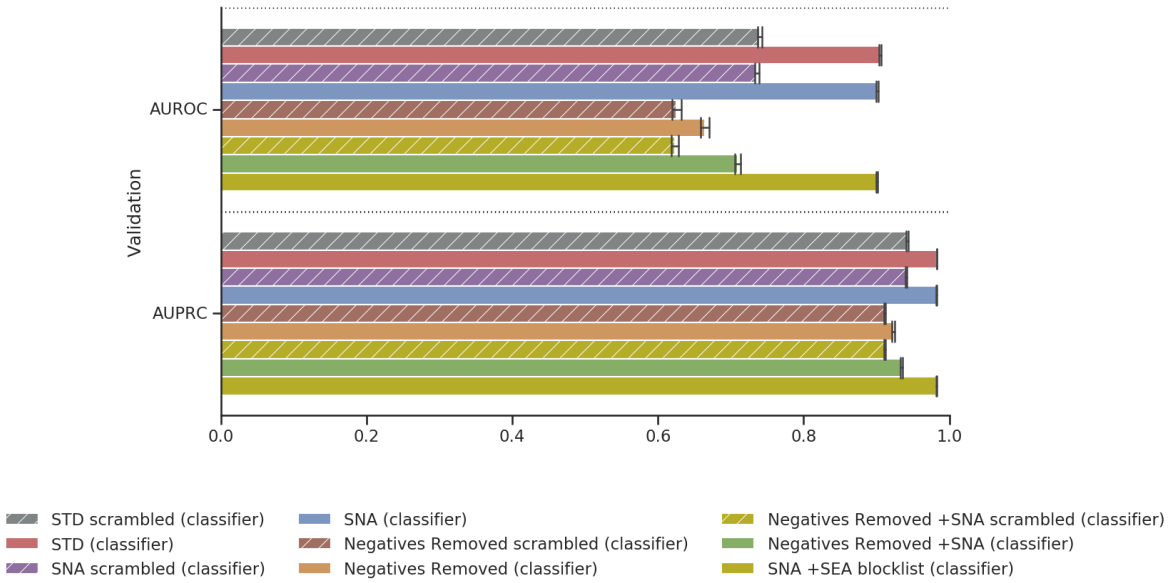

Figure A.7: Validation performance for all classification models.

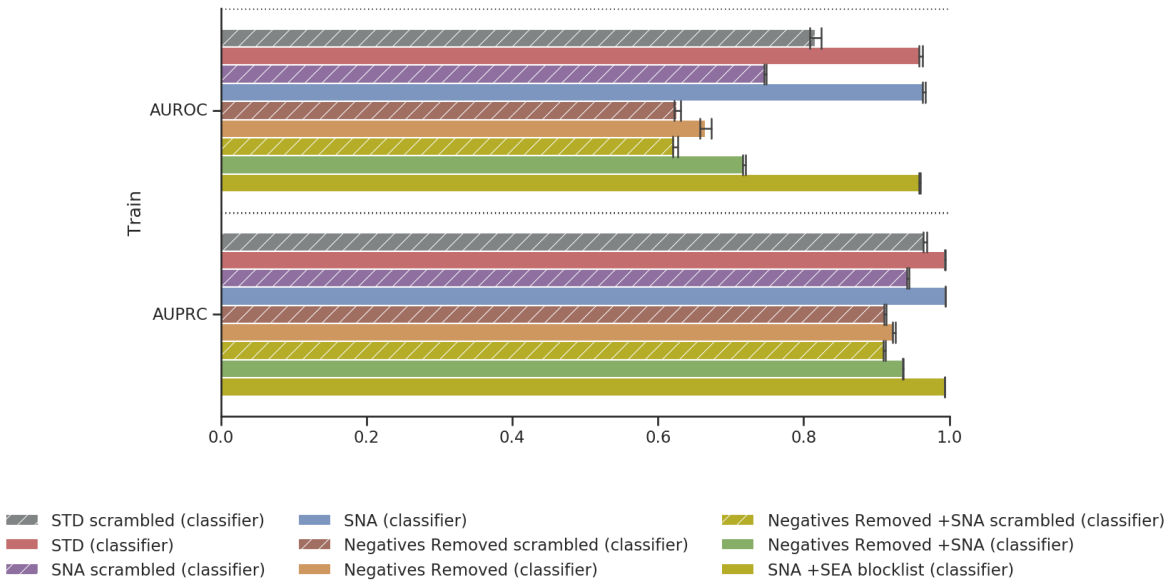

Figure A.8: Train performance for all classification models.

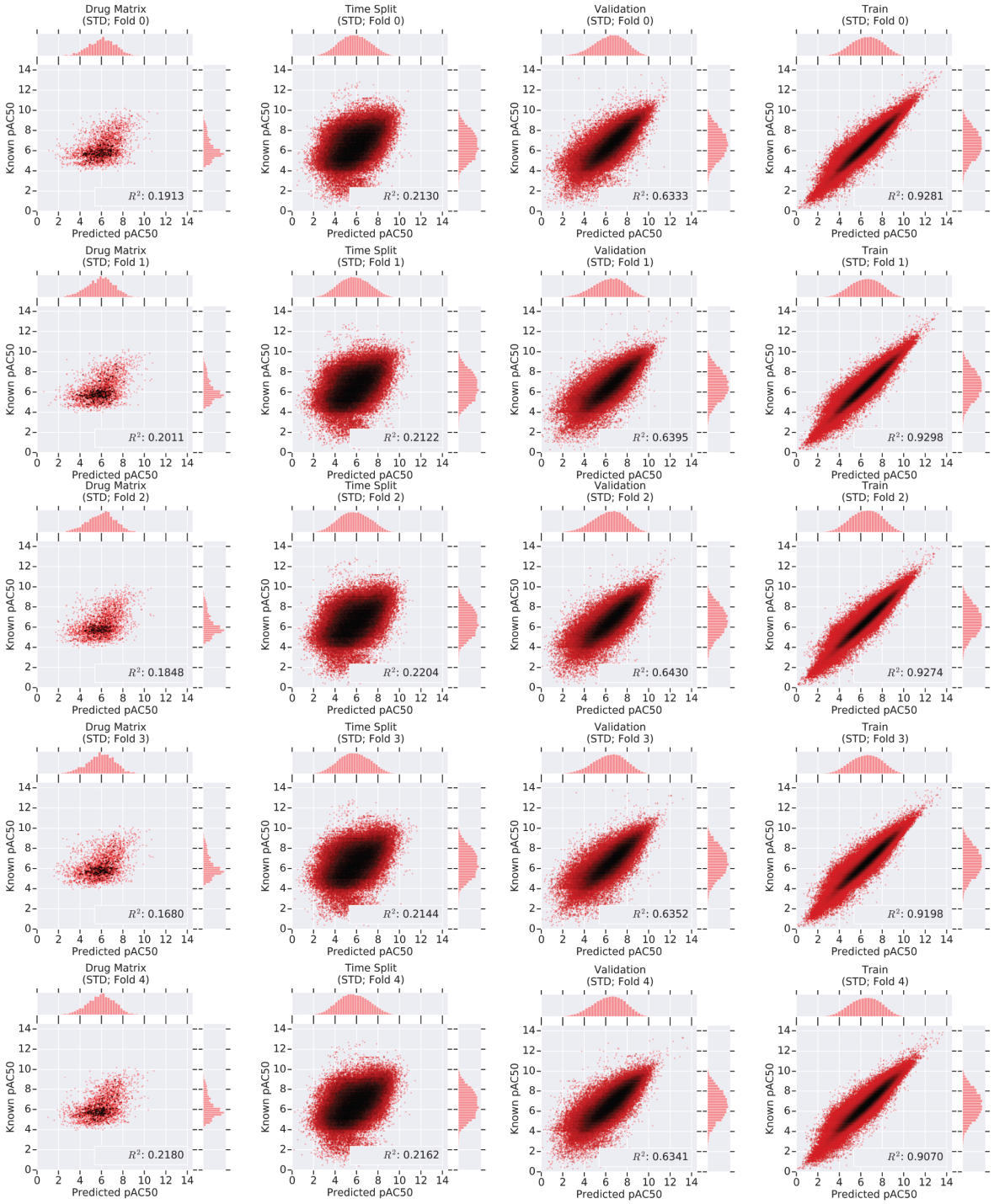

Figure A.9: STD model R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation(column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing).

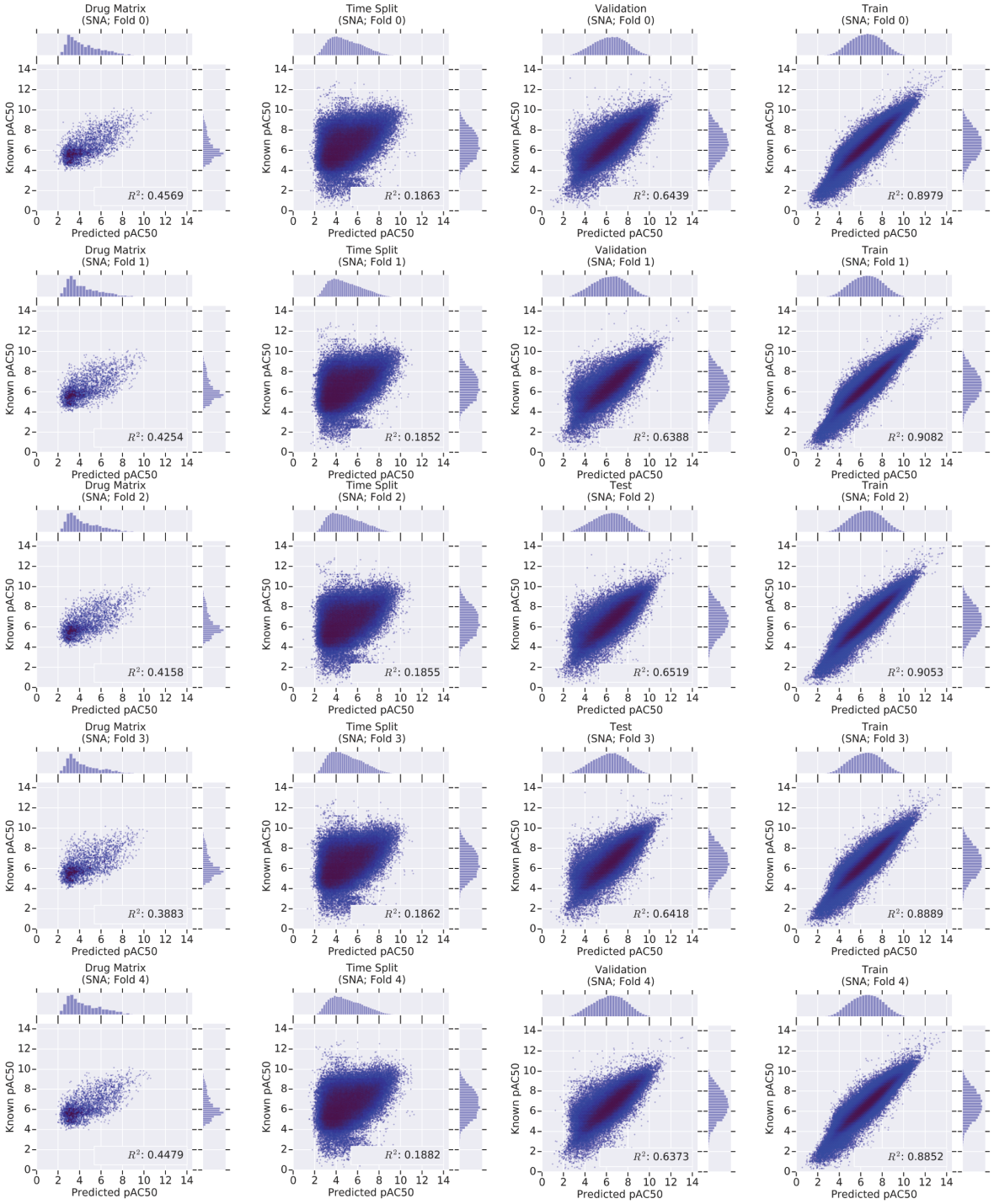

Figure A.10: SNA model R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation (column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing).

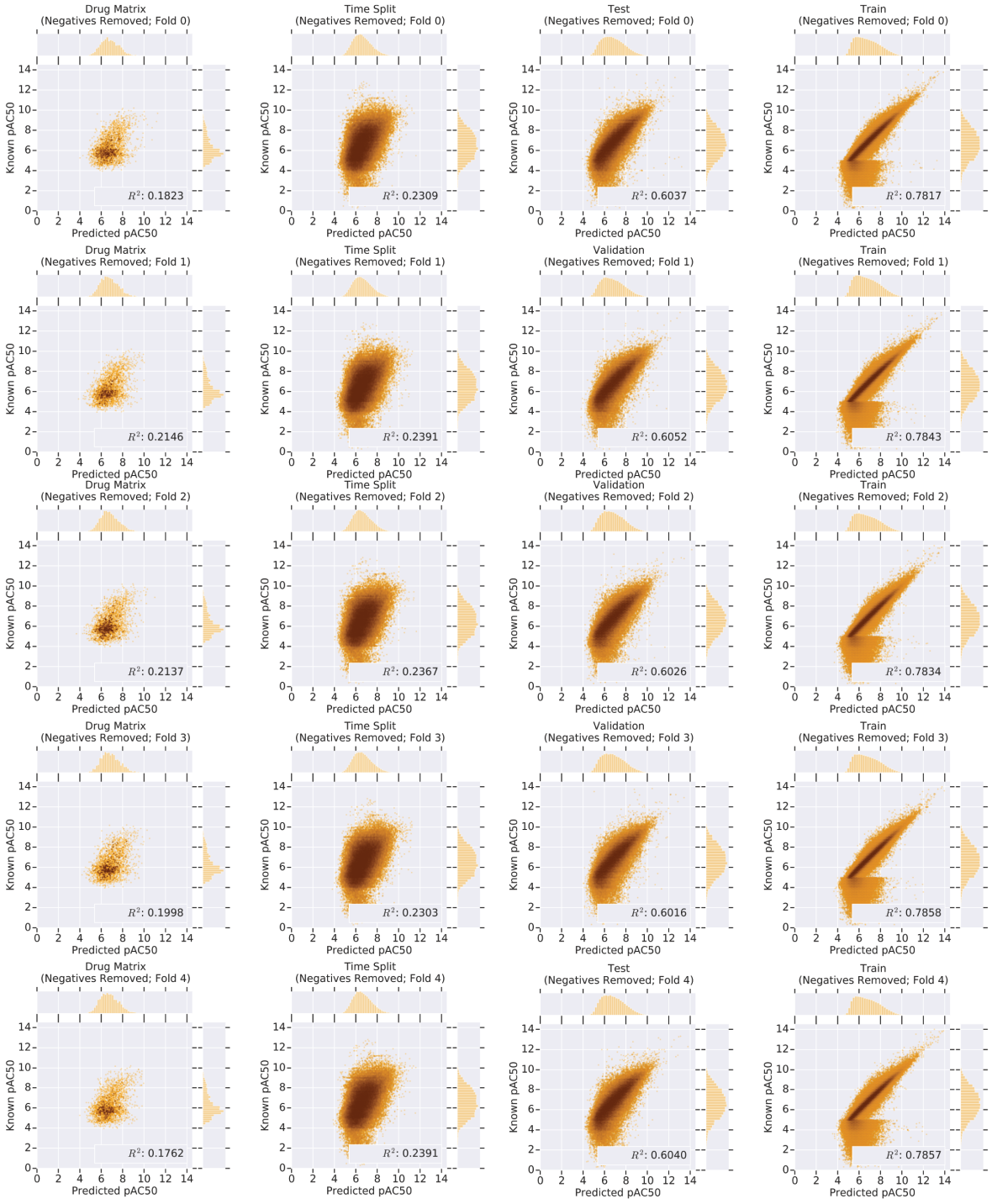

Figure A.11: Negatives Removed DNN R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation(column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing).

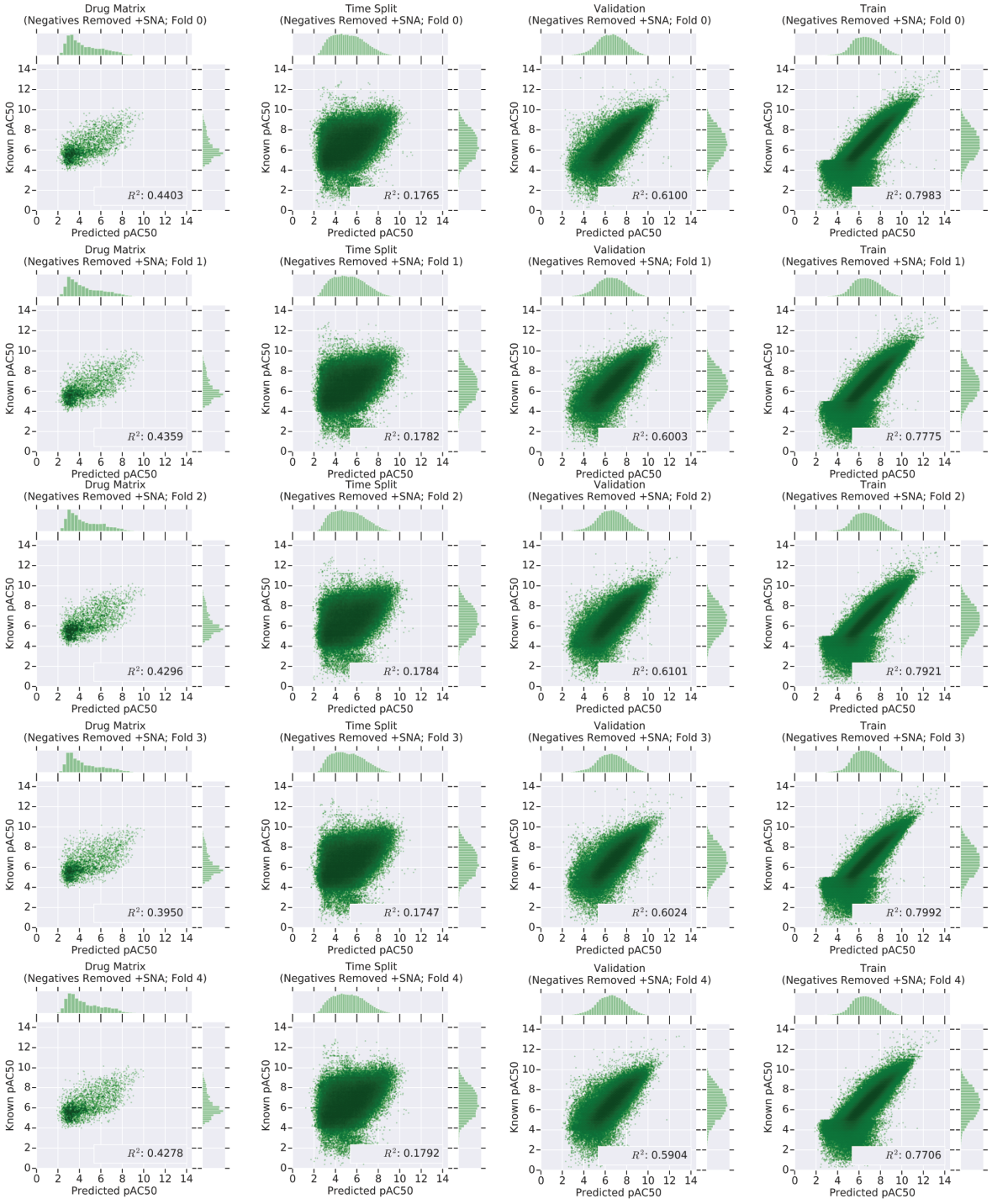

Figure A.12: Negatives Removed +SNA DNN R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation (column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing).

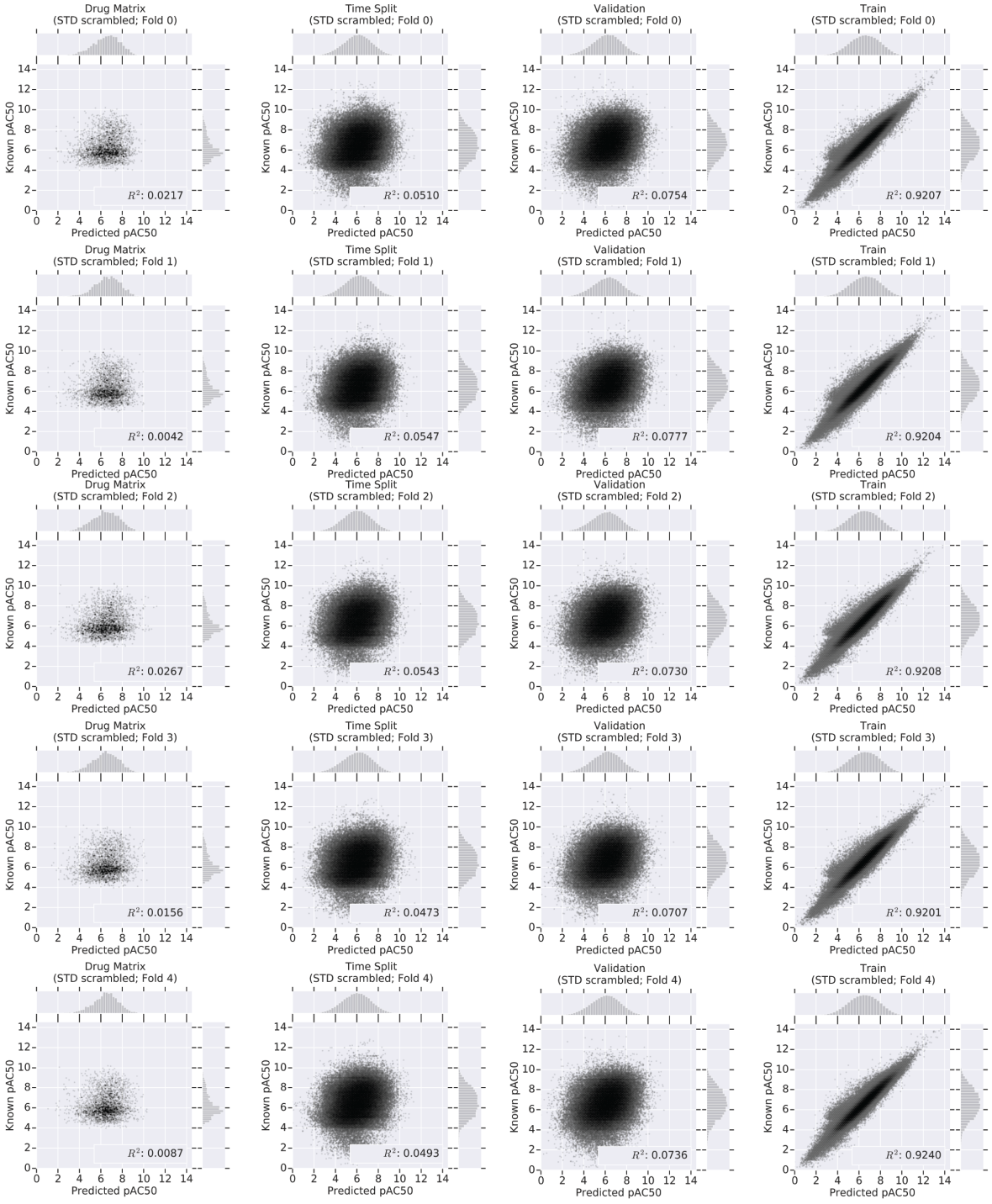

Figure A.13: STD scrambled (y-randomized training set control) DNN R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation (column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing).

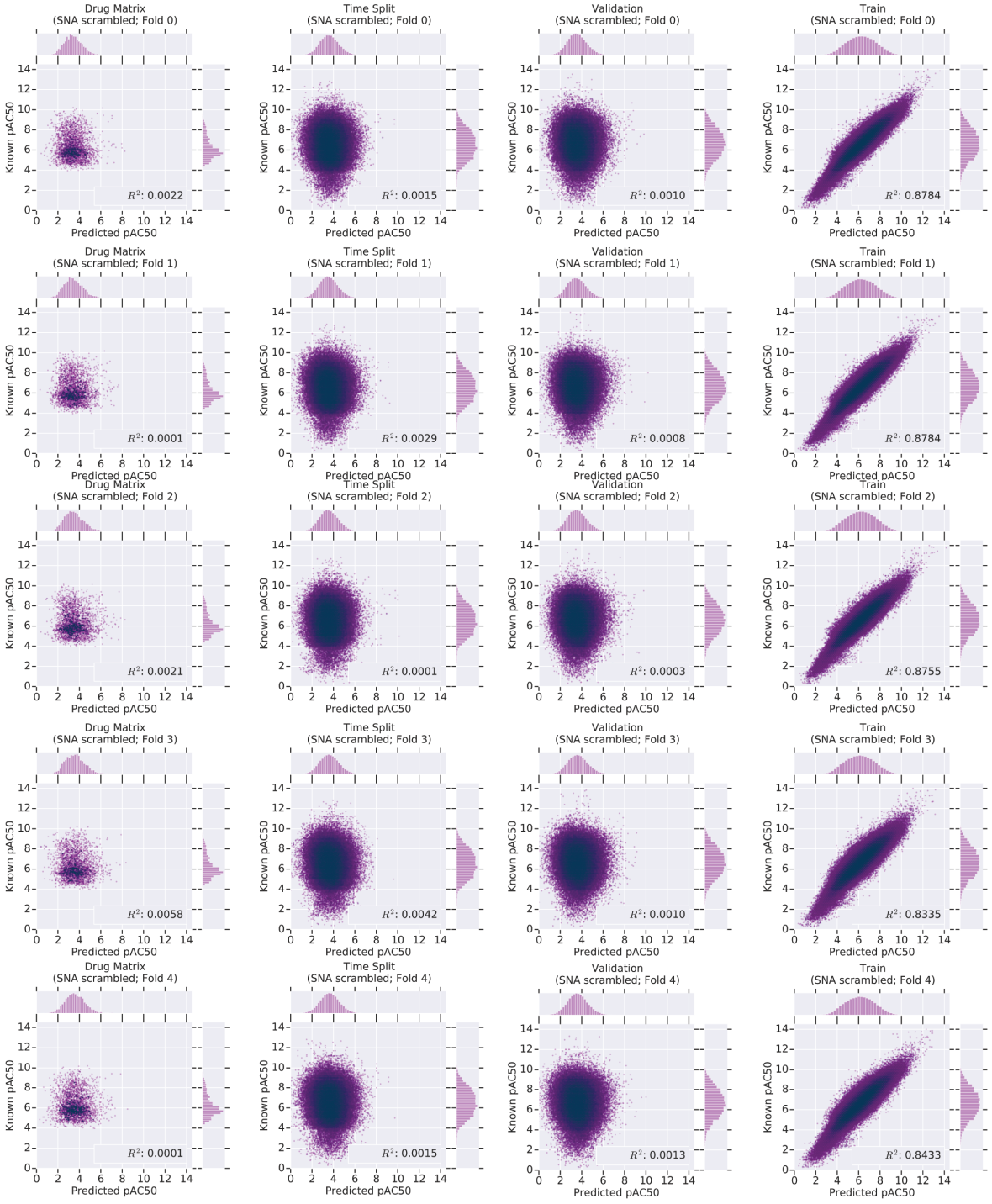

Figure A.14: SNA scrambled (y-randomized training set control with stochastic negatives) DNN R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation (column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing).

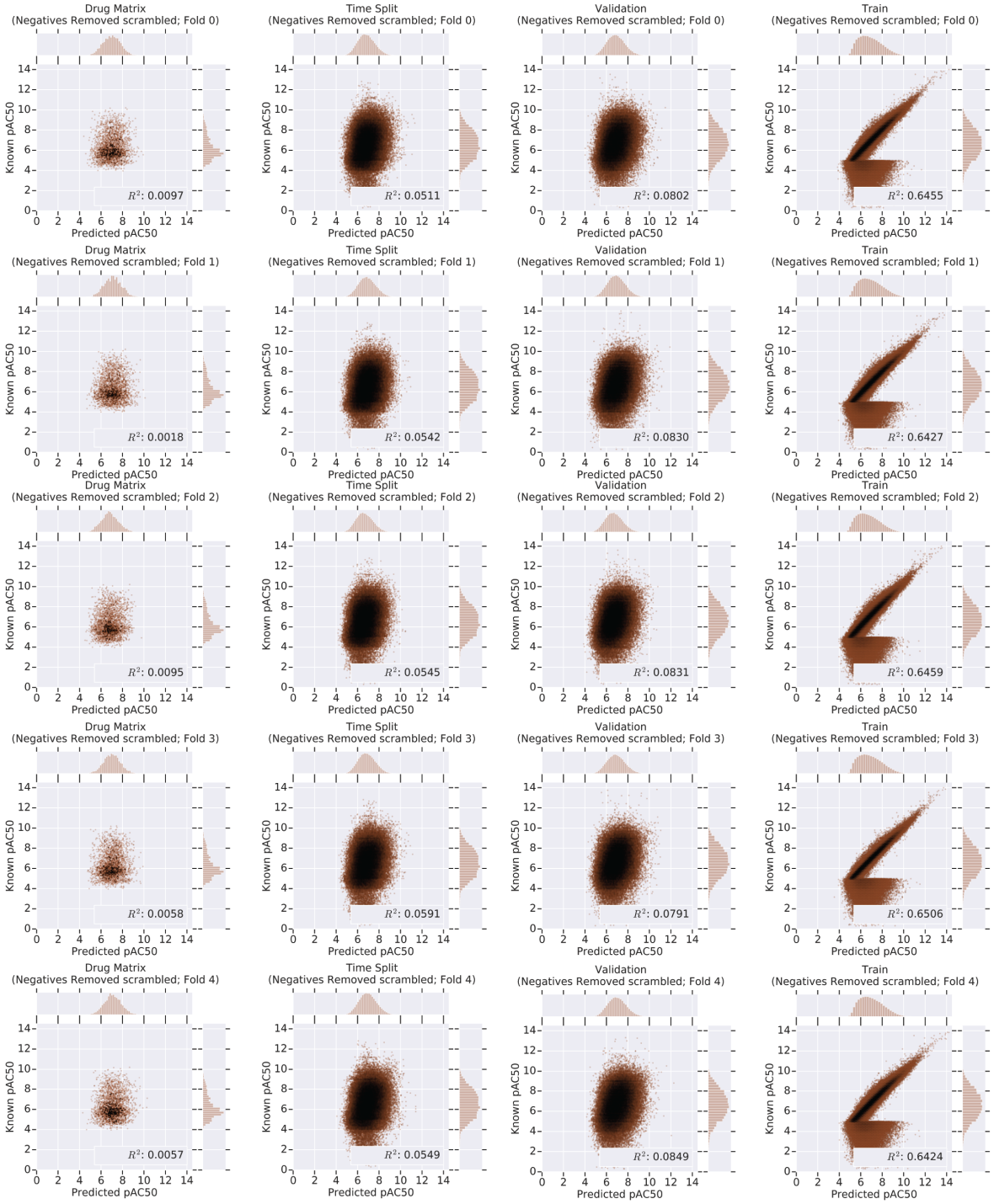

Figure A.15: Negatives Removed scrambled (y-randomized training set control with Negatives removed from the training set) DNN R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation (column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing). Note that "Train" (rightmost column) plots show a discontinuity at y < 5 because this is a Negatives Removed scenario, such that these negative examples were removed during model training, but nonetheless retained and used for model evaluation in this plot.

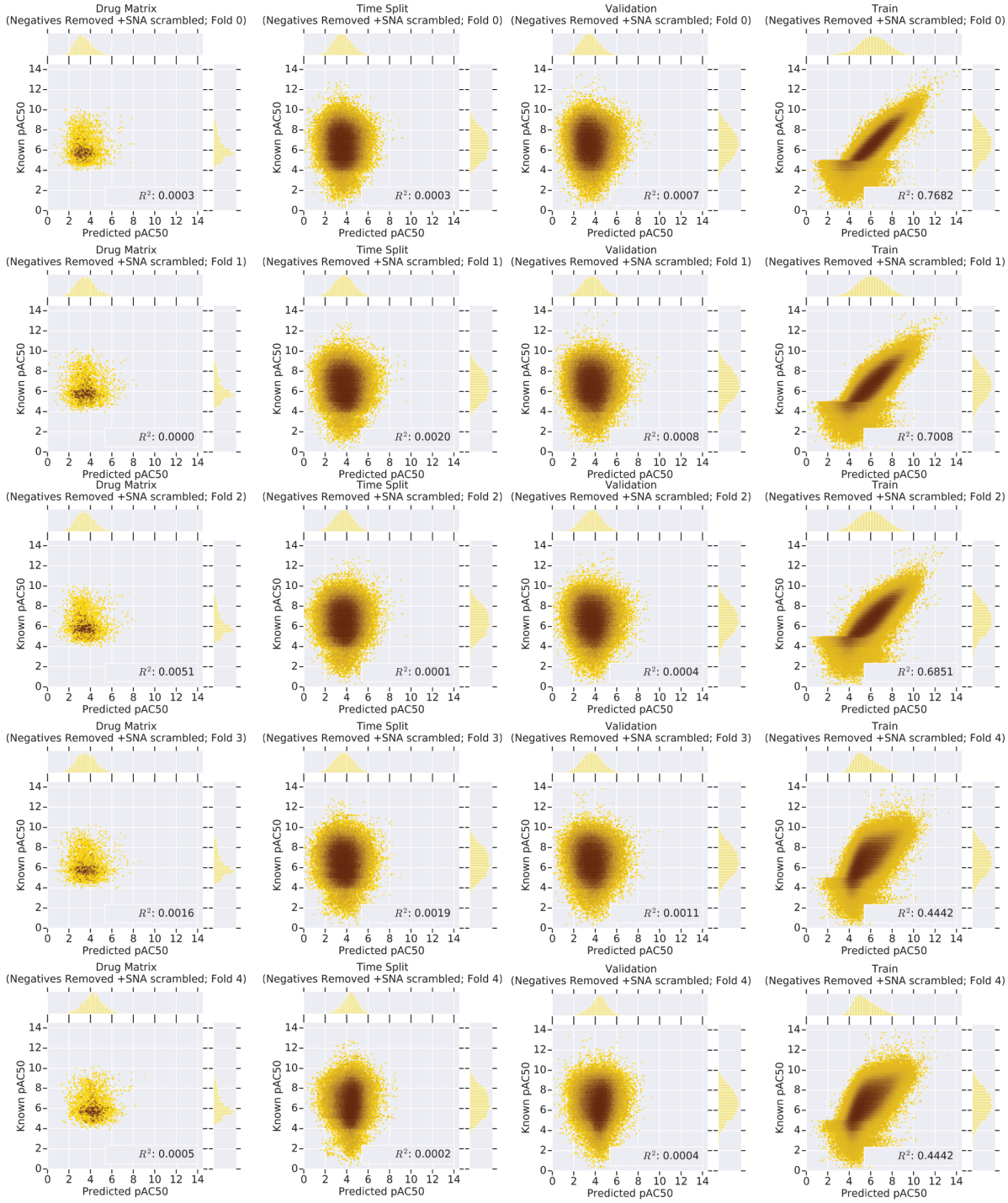

Figure A.16: Negatives Removed +SNA scrambled (y-randomized training set control with stochastic negatives) DNN R^2 plots for Drug Matrix (column 1), Time Split (column 2), Validation (column 3), and Train (column 4) across each fold (0-4, top to bottom, increasing). Note that "Train" (rightmost column) plots show a discontinuity at $y < 5$ (although less than that of Figure S15) because this is a Negatives Removed scenario, such that these negative examples were removed during model training, but nonetheless retained and used for model evaluation in this plot.

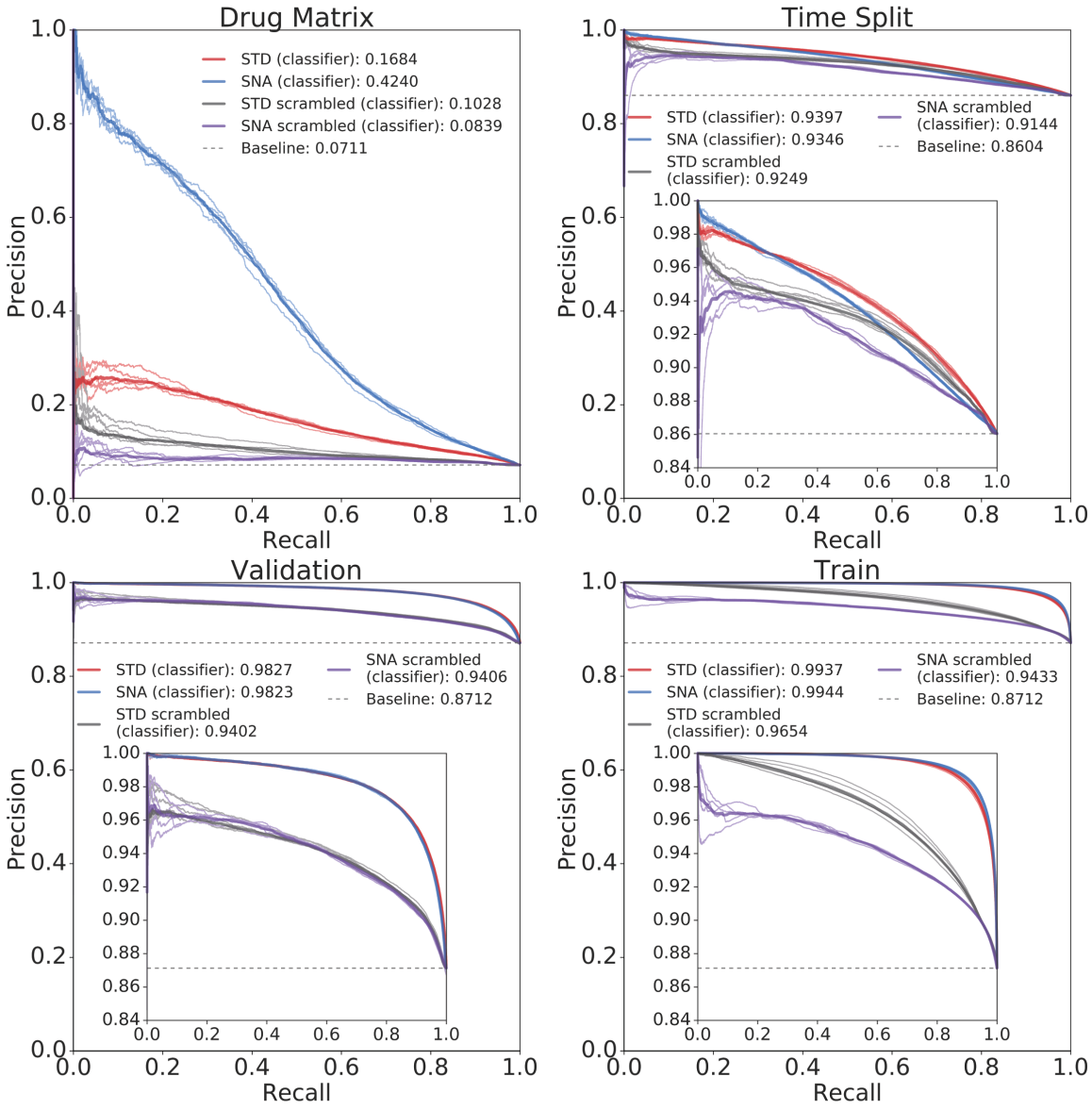

Figure A.17: AUPRC plots for SNA, STD, SNA scrambled, and STD scrambled classification DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUPRC plotted with a thicker line.

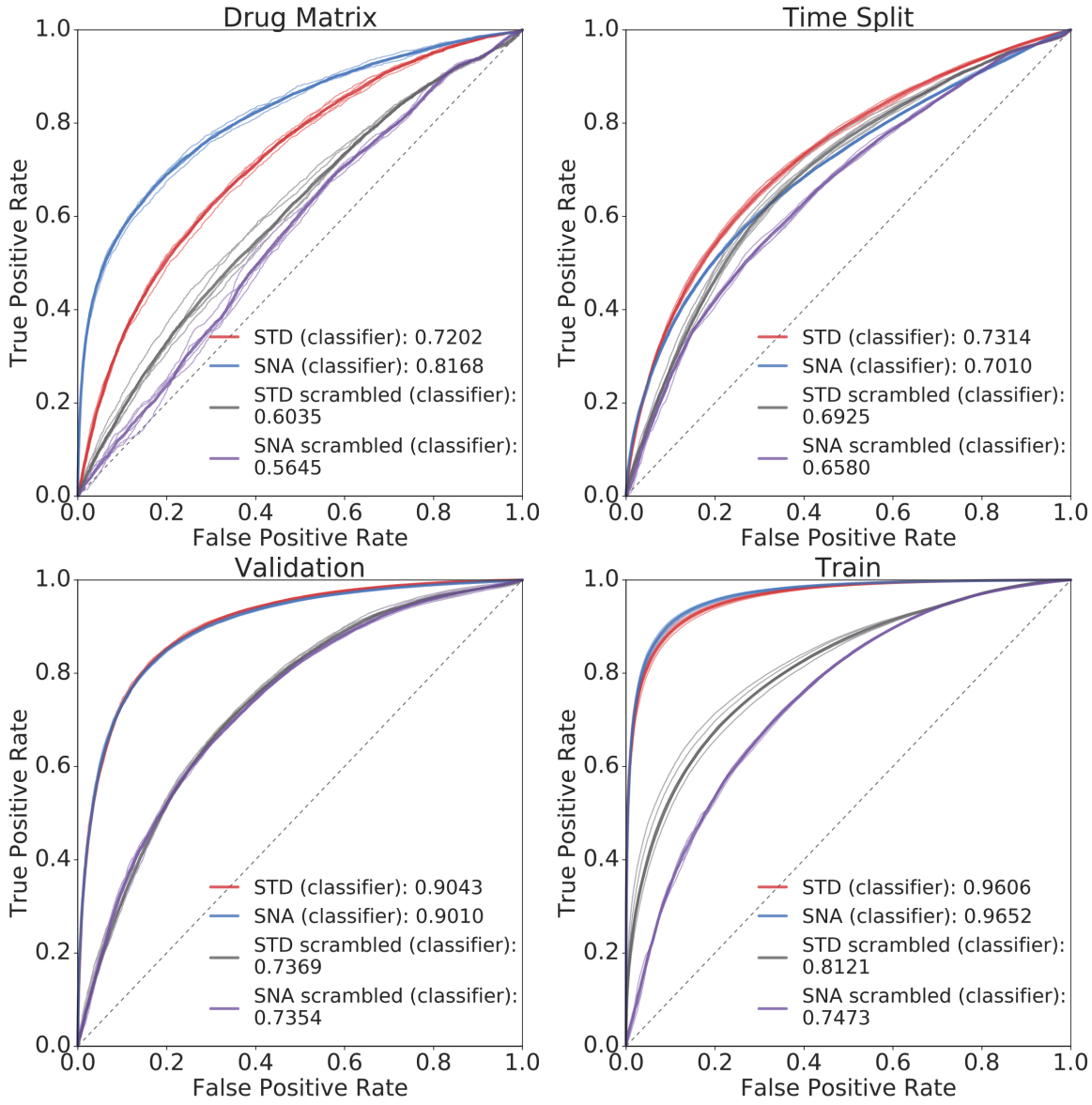

Figure A.18: AUROC plots for SNA, STD, SNA scrambled, and STD scrambled classification DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC plotted with a thicker line.

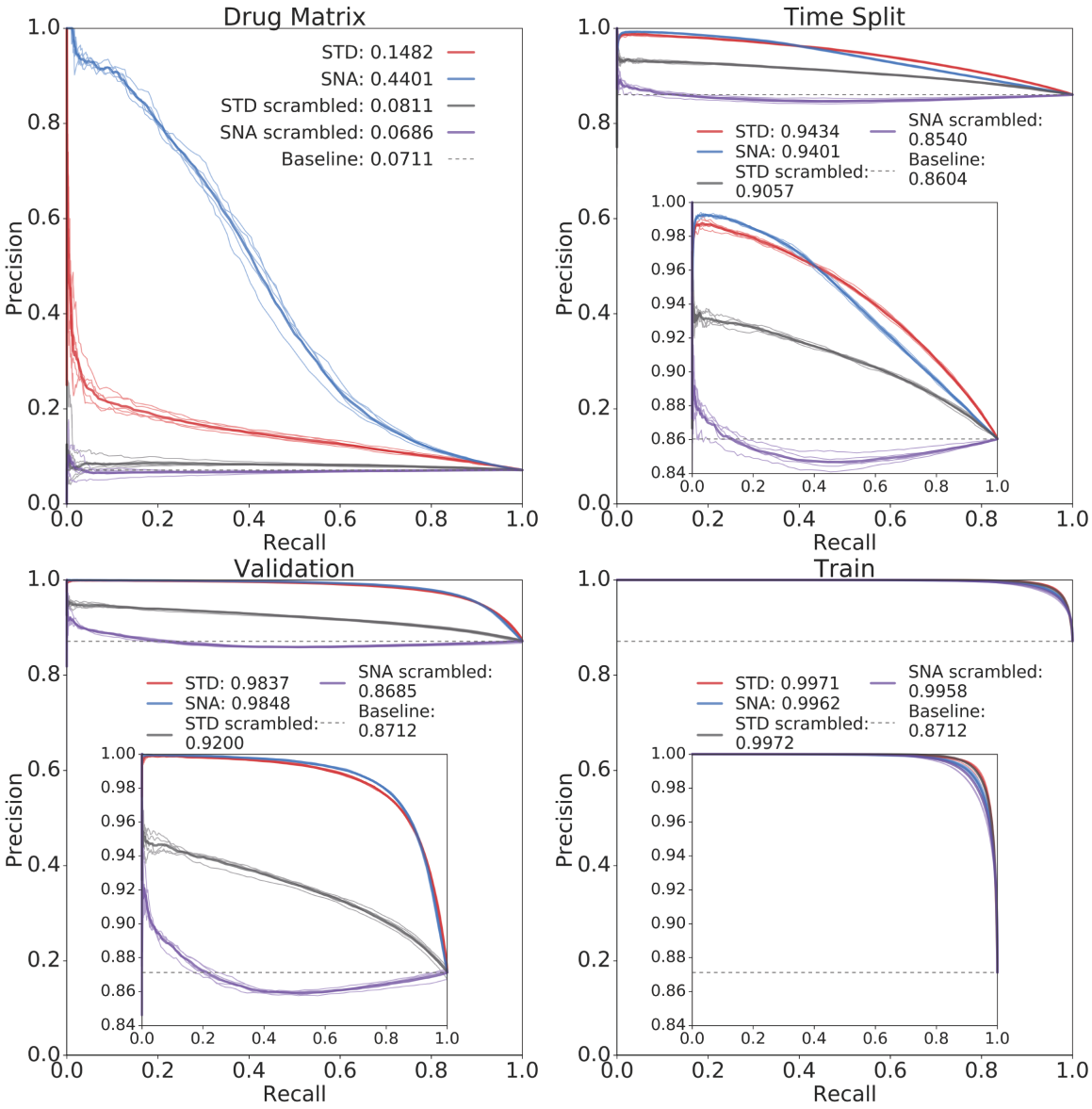

Figure A.19: AUPRC_r plots for SNA, STD, SNA scrambled, and STD scrambled regression DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUPRC_r plotted with a thicker line.

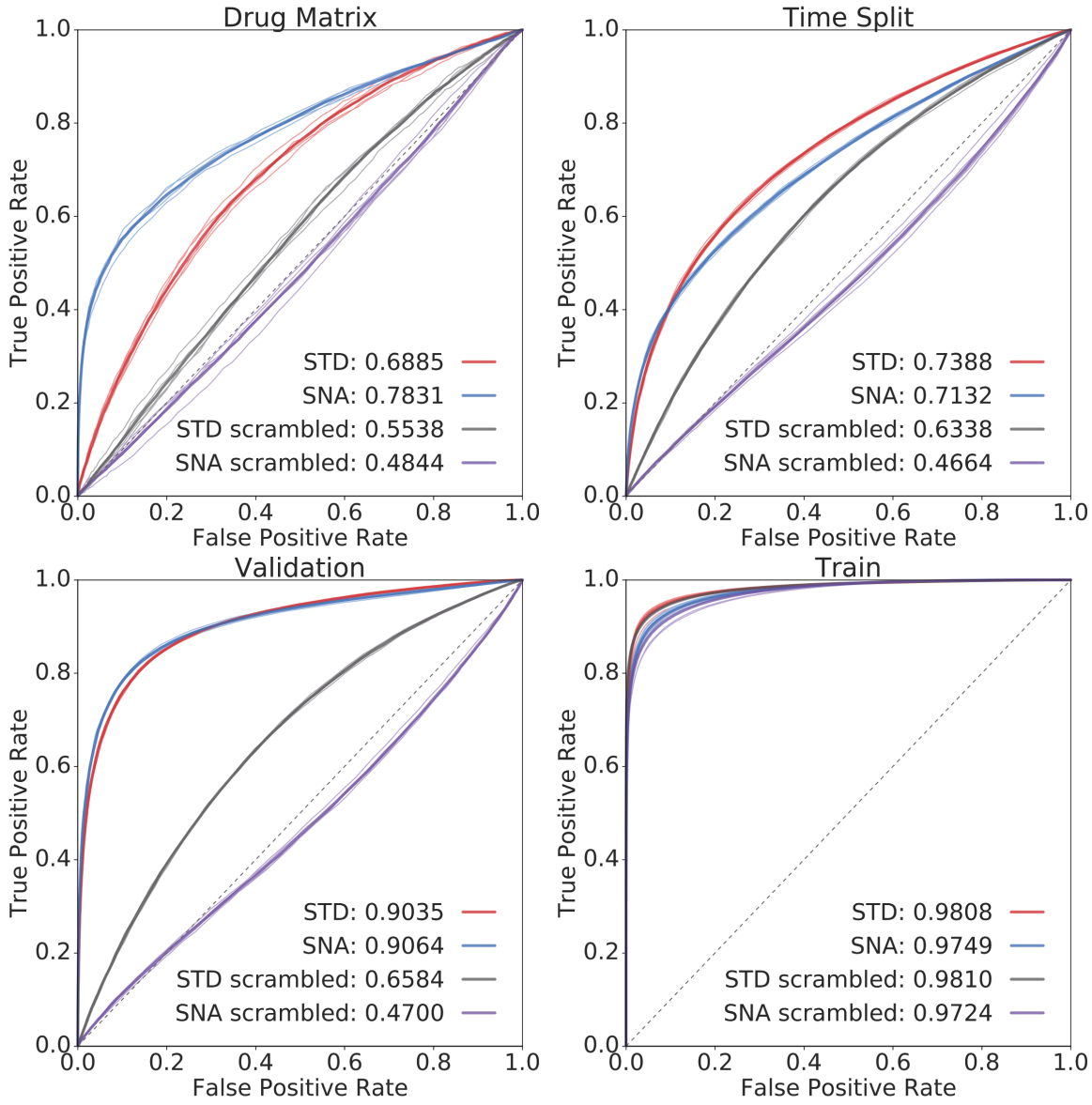

Figure A.20: AUROC_r plots for SNA, STD, SNA scrambled, and STD scrambled regression DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC_r plotted with a thicker line.

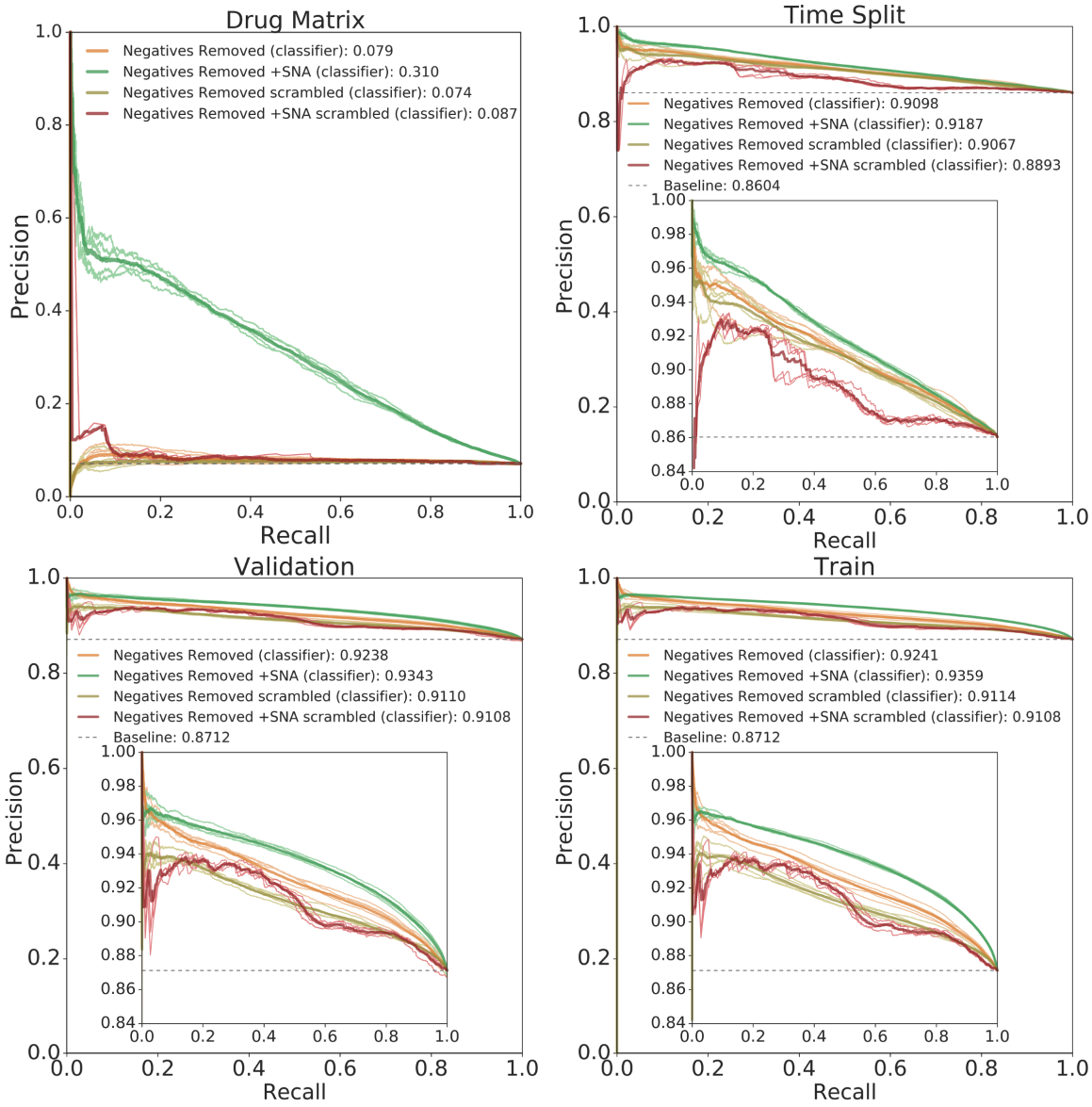

Figure A.21: AUPRC plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled classification DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUPRC plotted with a thicker line.

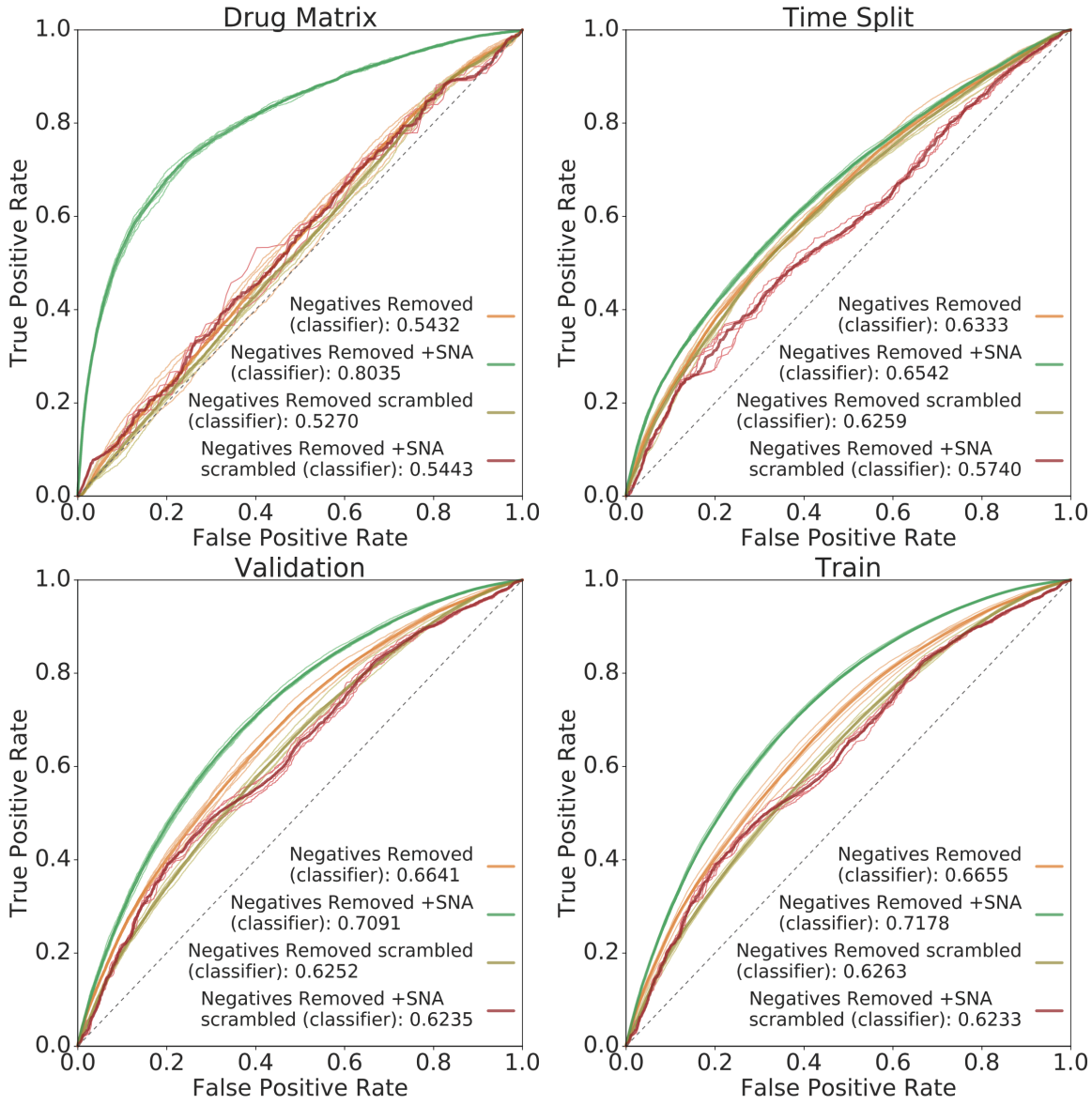

Figure A.22: AUROC plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled classification DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC plotted with a thicker line.

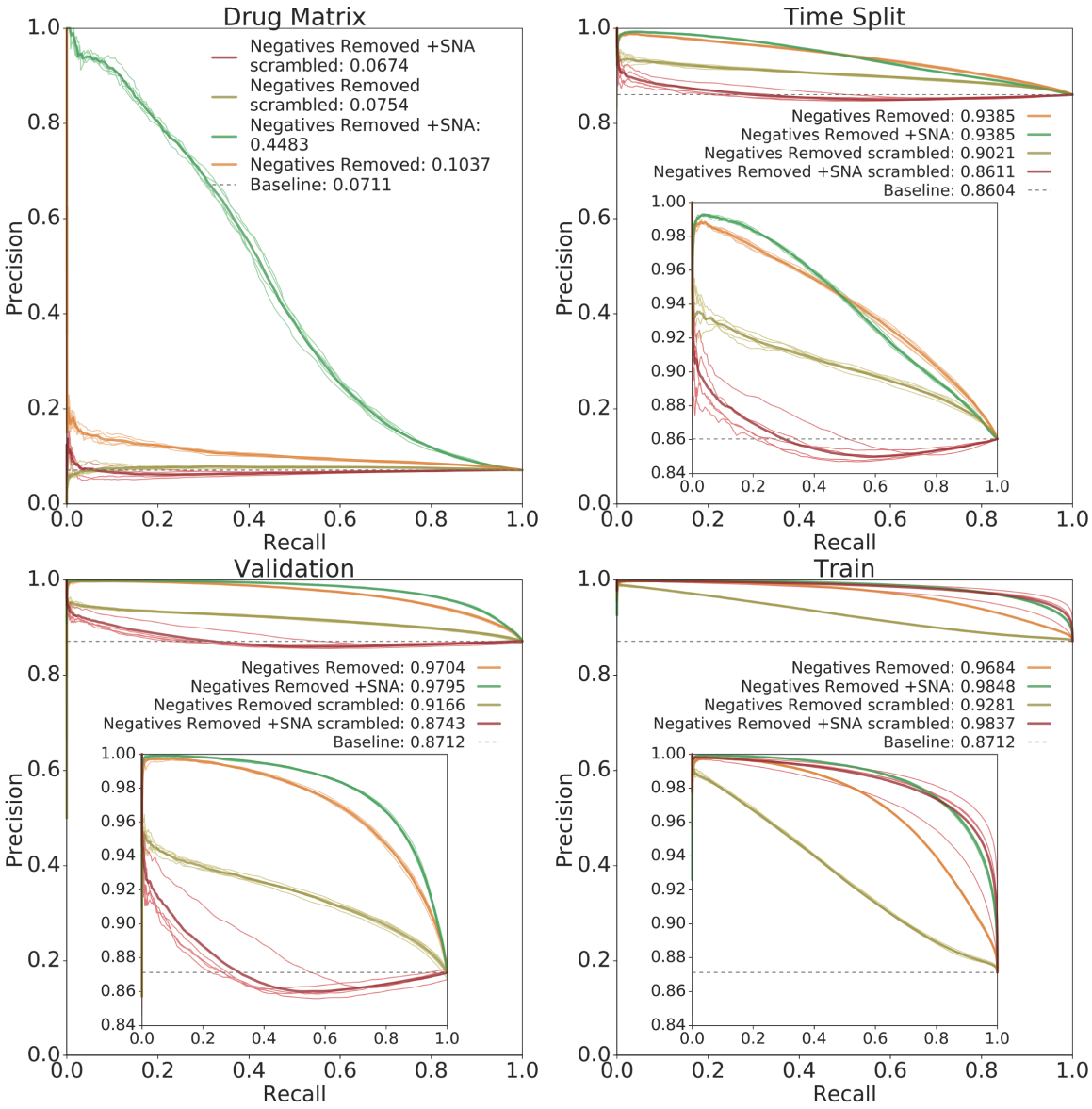

Figure A.23: AUPRC_r plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled regression DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUPRC_r plotted with a thicker line.

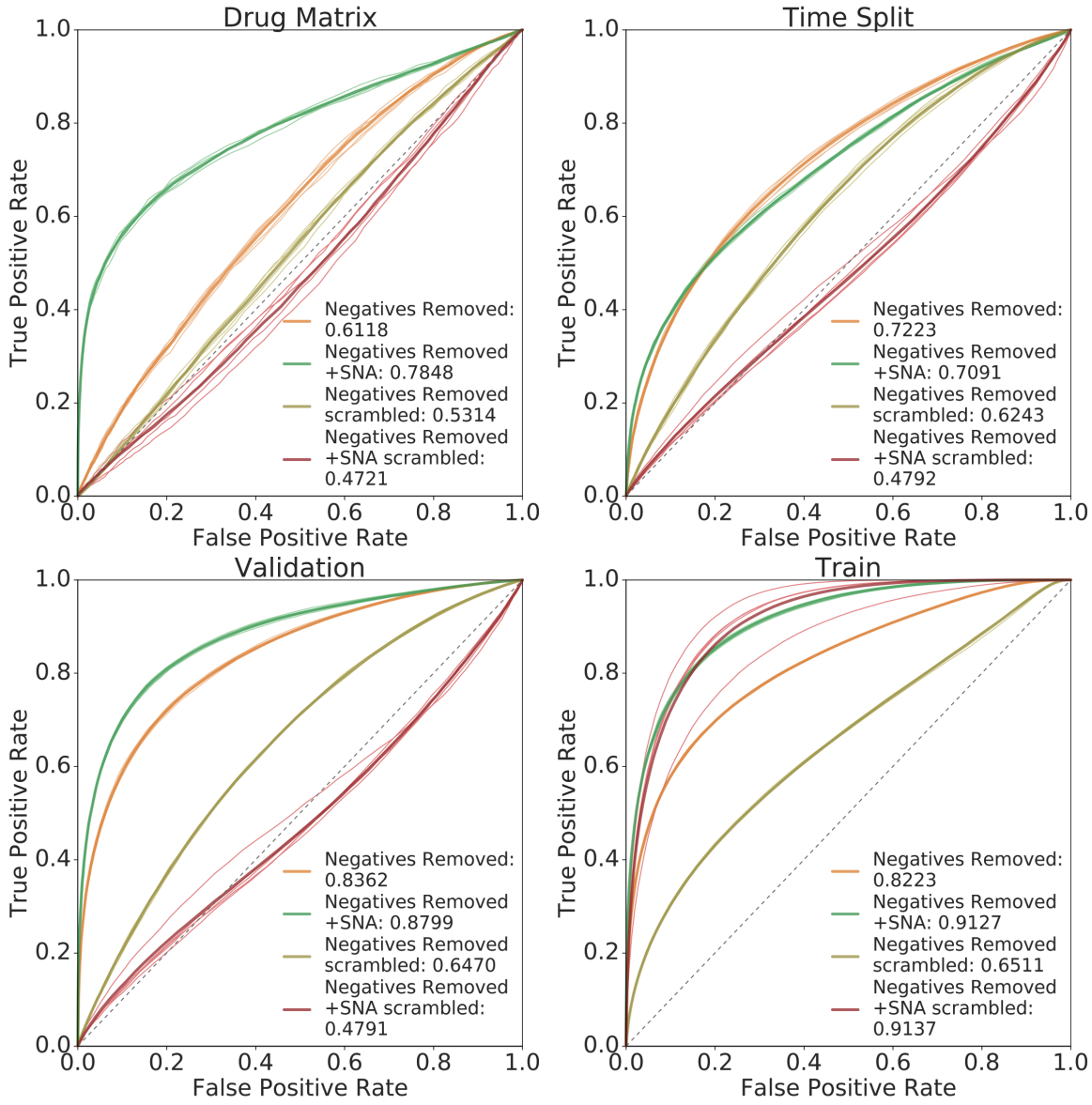

Figure A.24: AUROC_r plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled regression DNNs for Drug Matrix (upper left), Time Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC_r plotted with a thicker line.

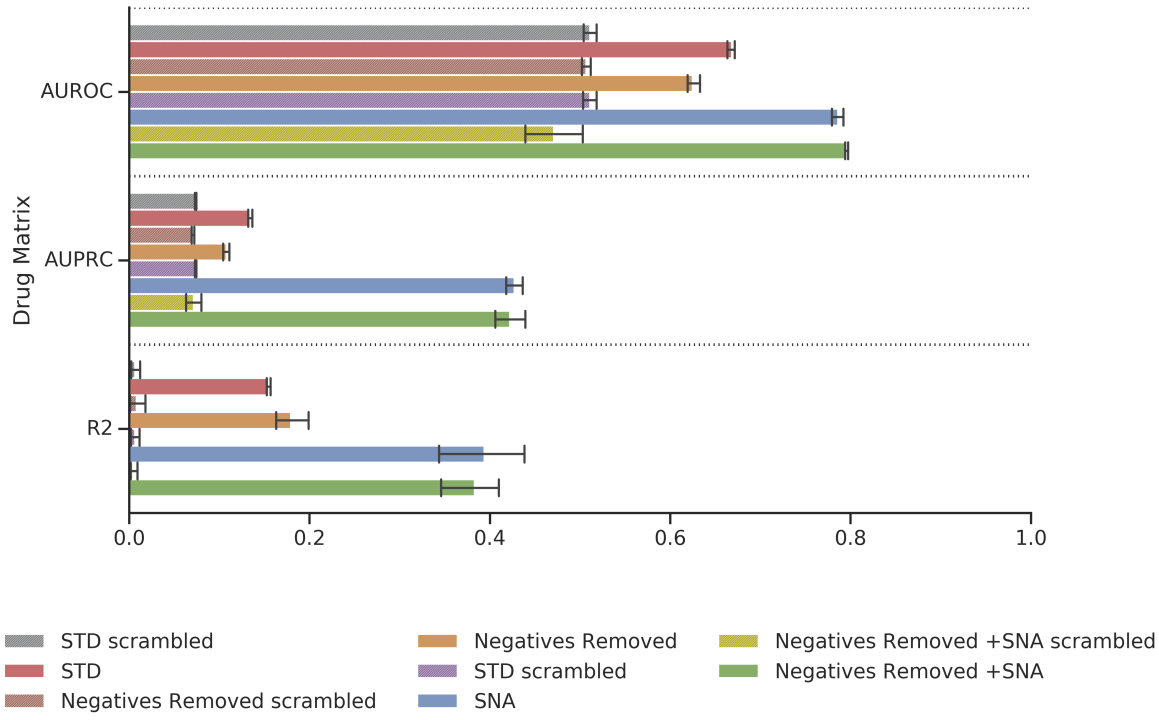

Figure A.25: Drug Matrix holdout performance for all Butina Split classification models.

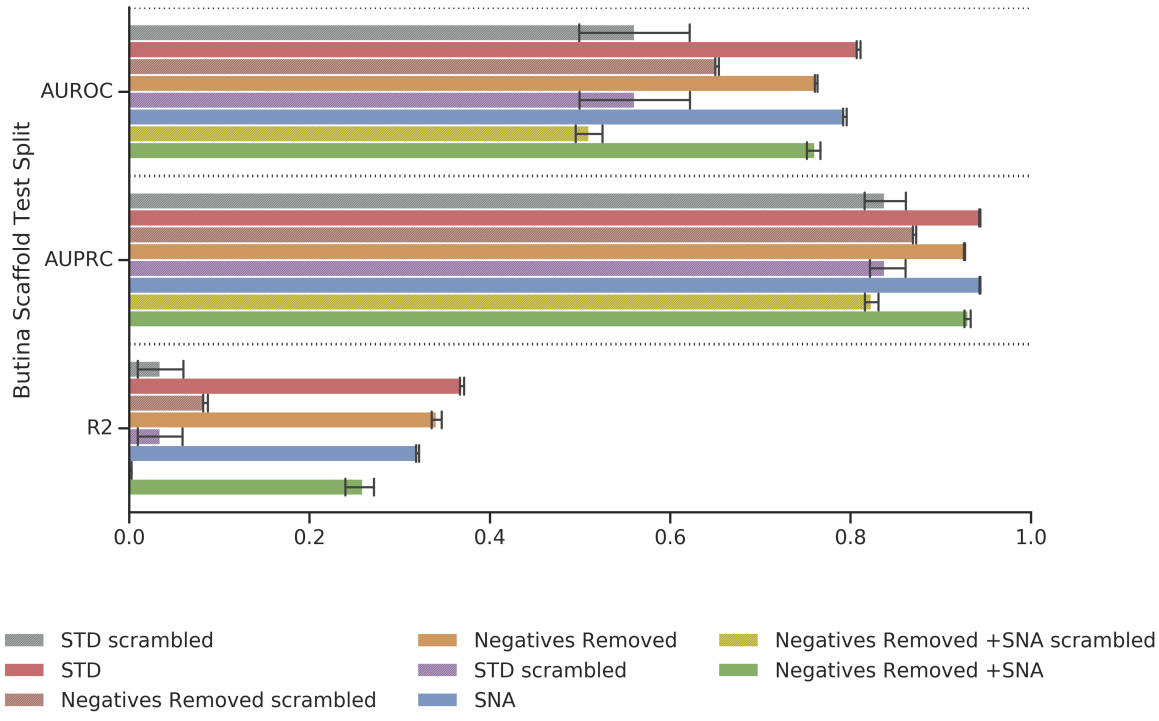

Figure A.26: Butina Scaffold Test Split holdout performance for all Butina Split classification models.

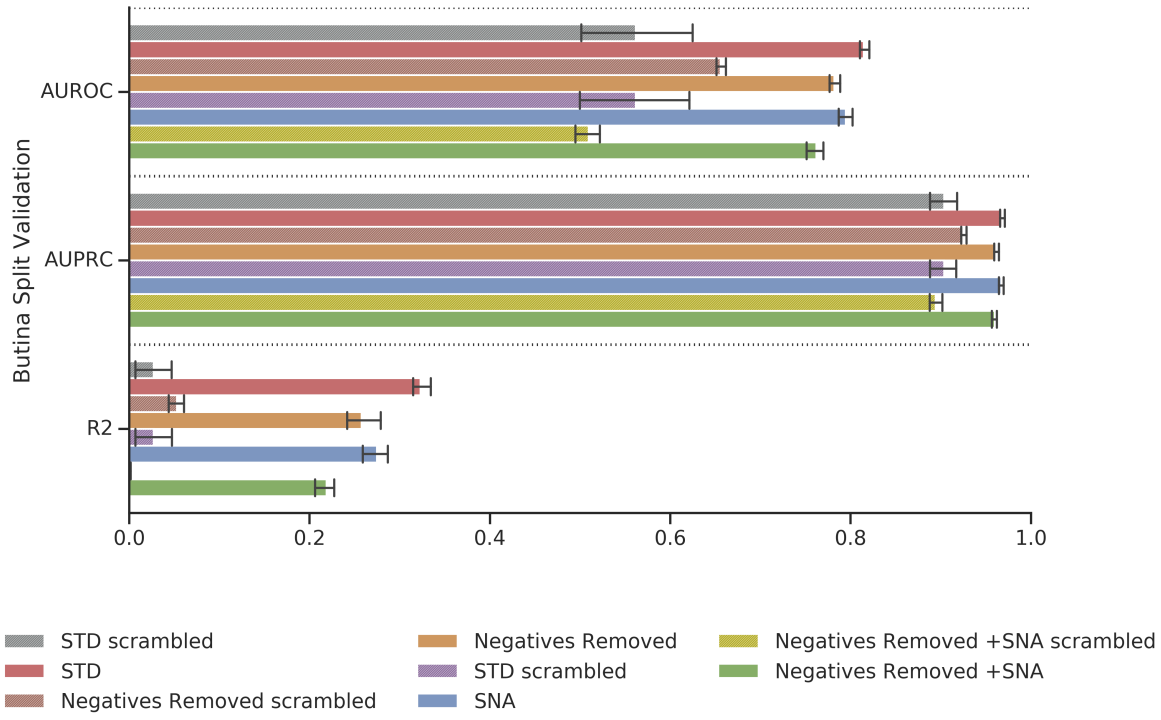

Figure A.27: Butina Split k-fold cross validation performance for all Butina Split classification models.

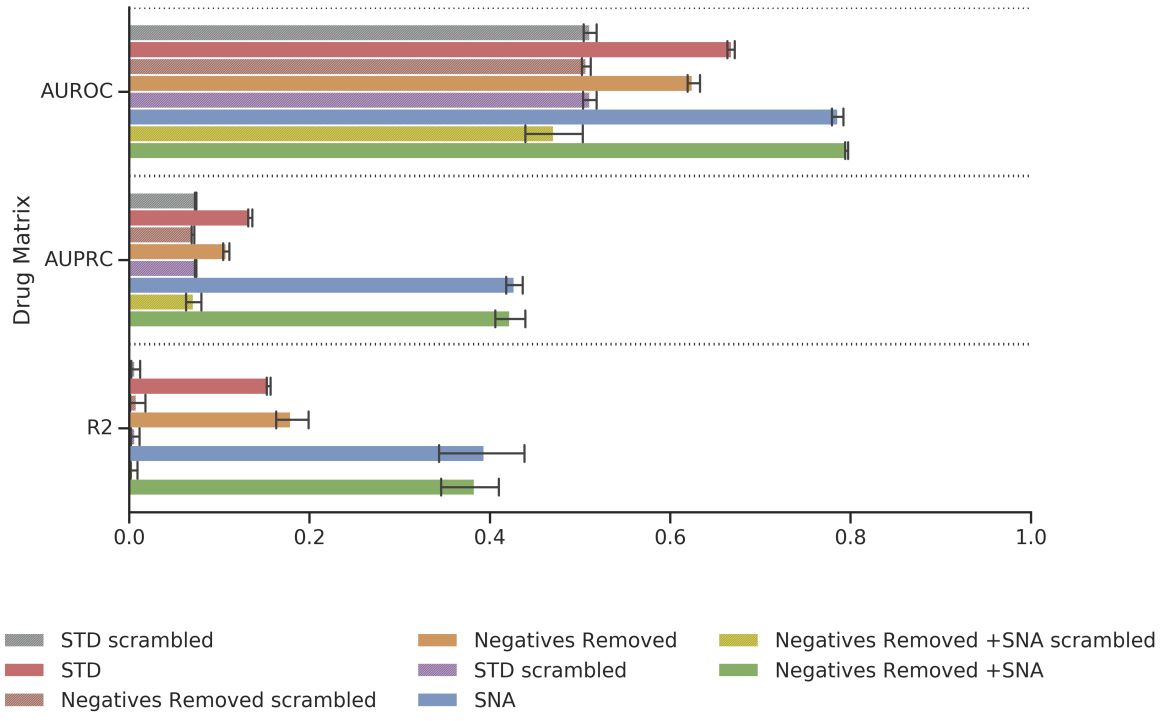

Figure A.28: Drug Matrix performance for all Butina Split regression models.

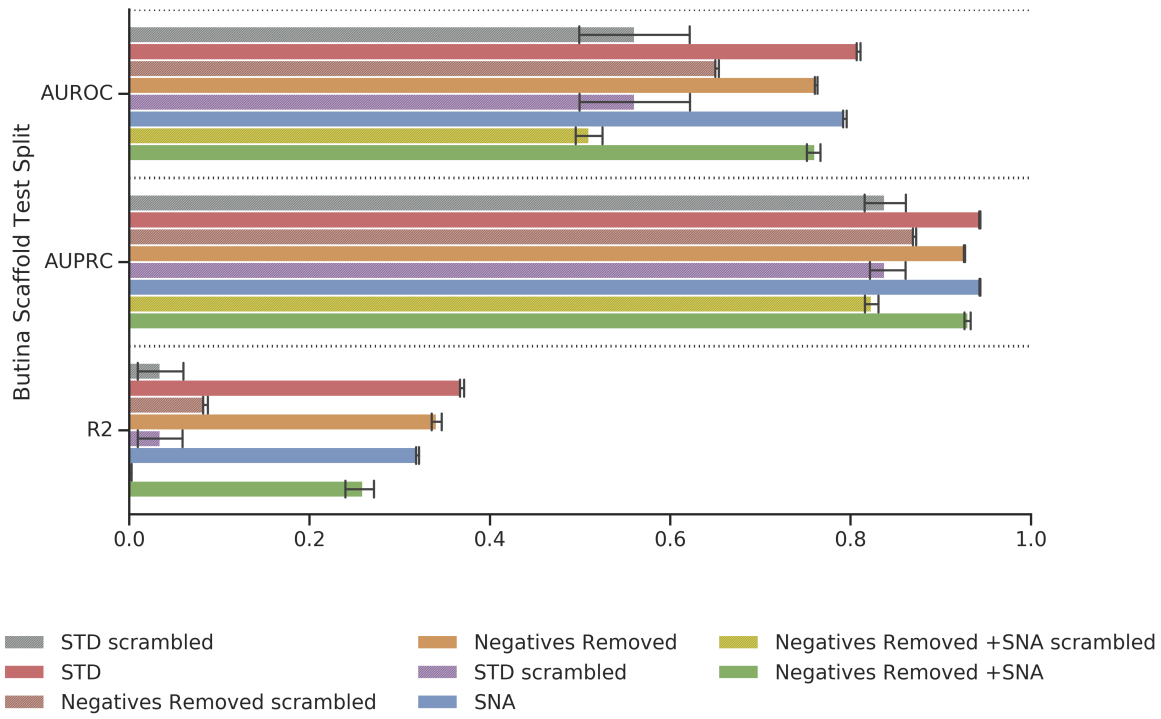

Figure A.29: Butina Scaffold Test Split holdout performance for all Butina Split regression models.

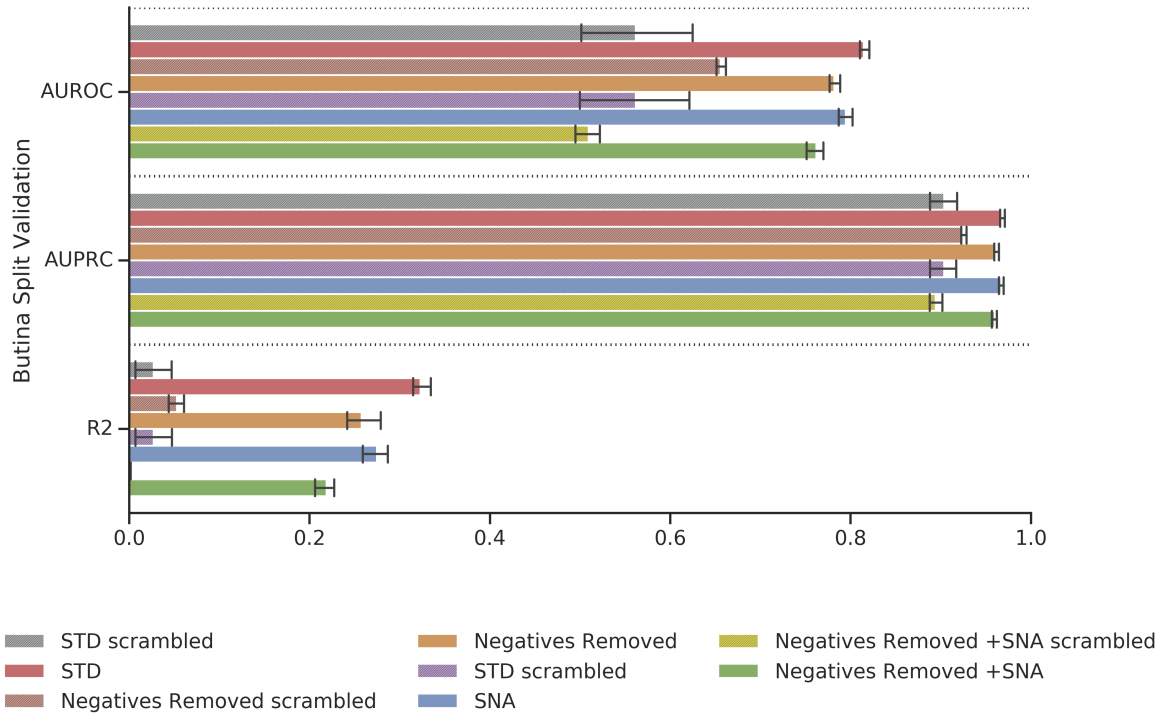

Figure A.30: Butina Split k-fold cross validation performance for all Butina Split regression models.

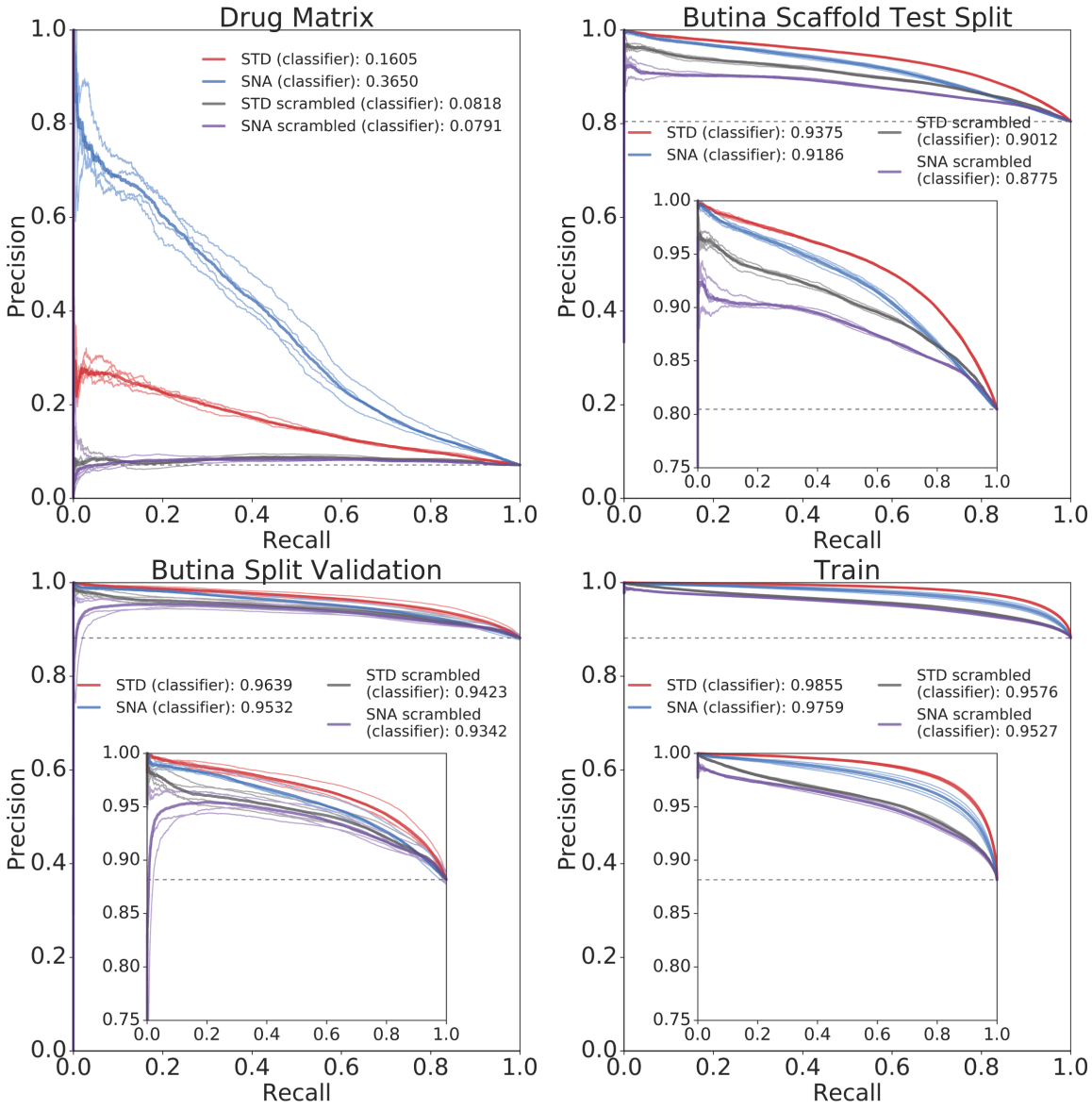

Figure A.31: Butina Split AUPRC plots for SNA, STD, SNA scrambled, and STD scrambled classification DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUPRC plotted with a thicker line.

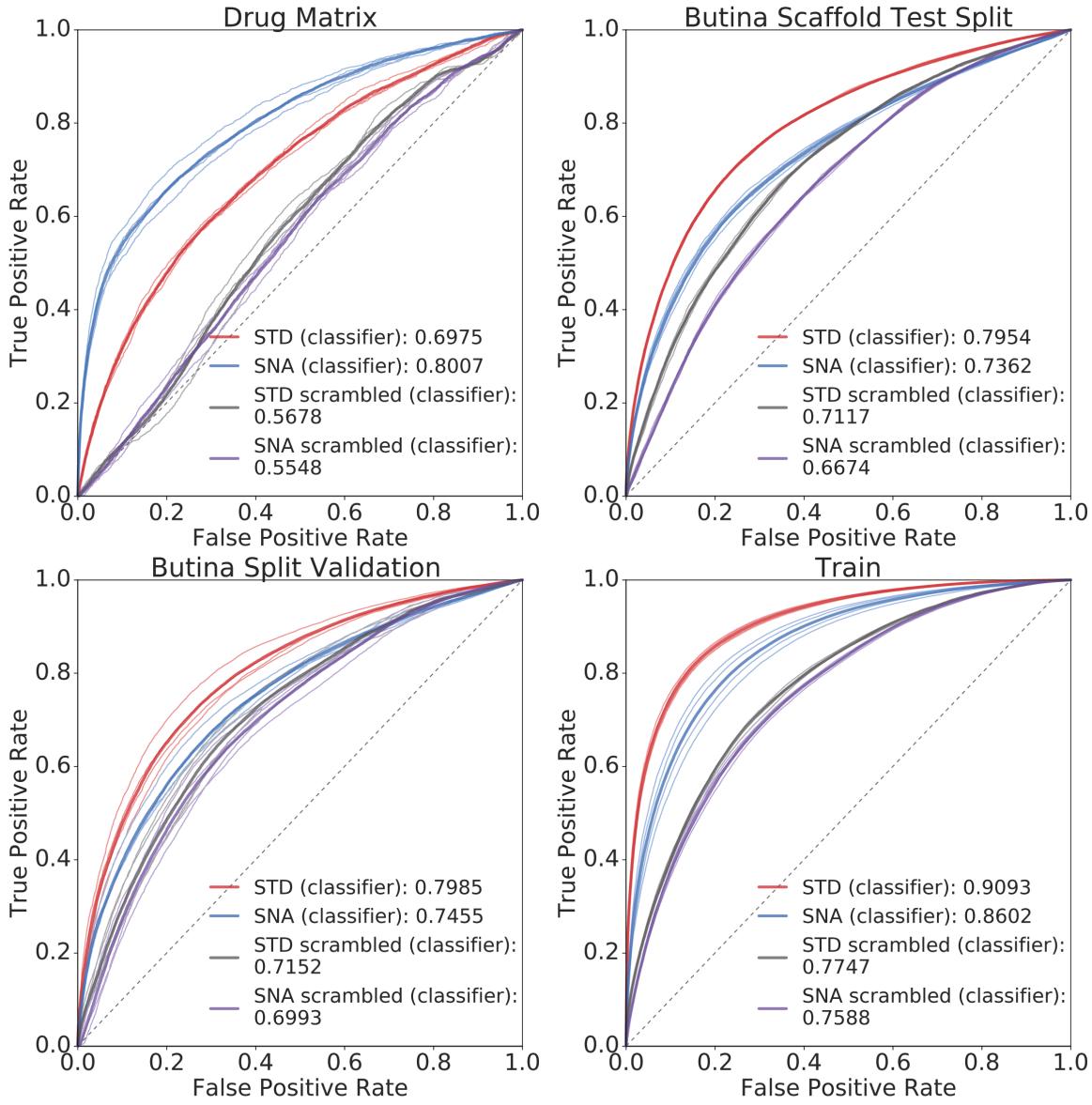

Figure A.32: Butina Split AUROC plots for SNA, STD, SNA scrambled, and STD scrambled classification DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC plotted with a thicker line.

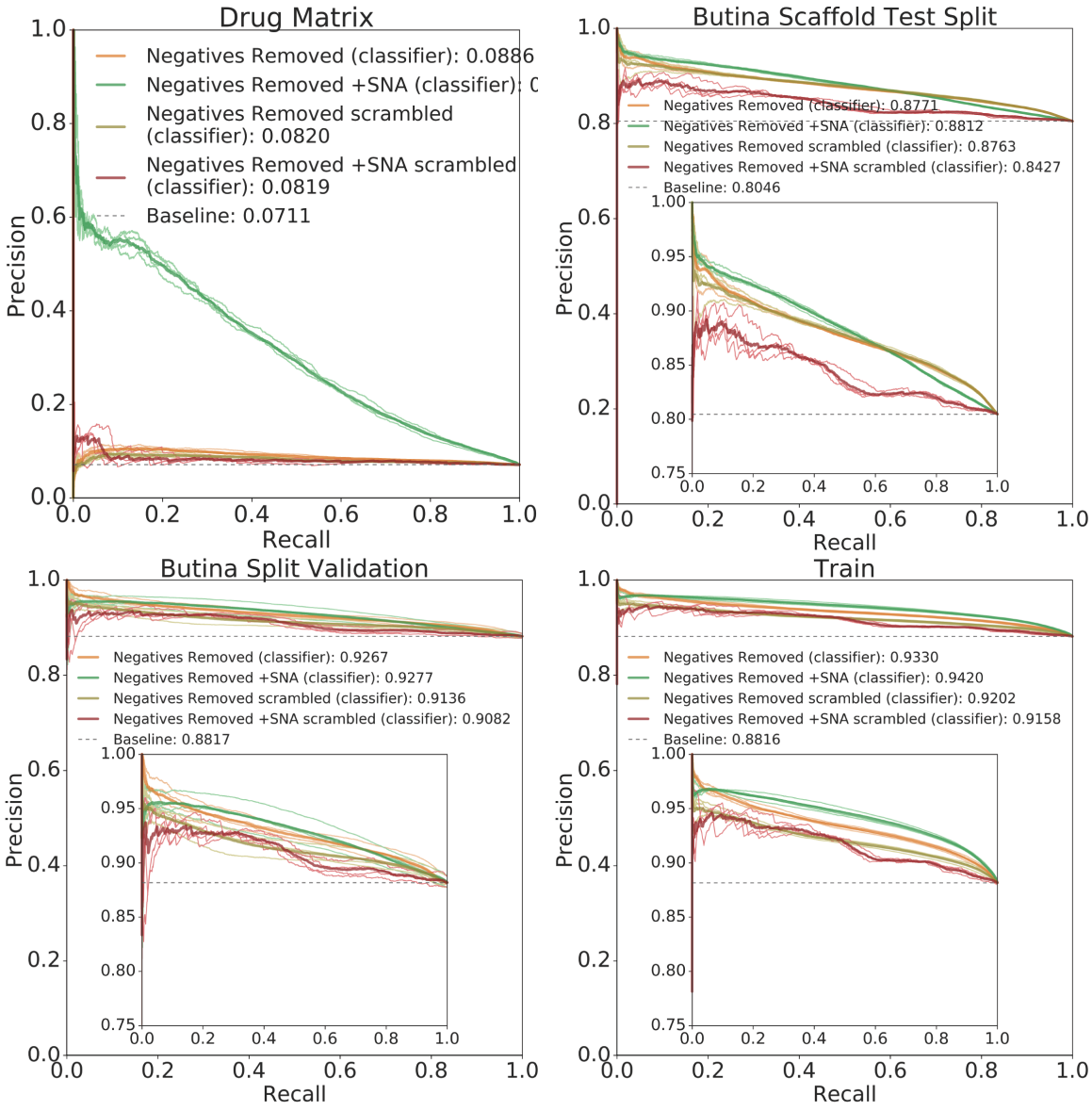

Figure A.33: Butina Split AUPRC plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled classification DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUPRC plotted with a thicker line.

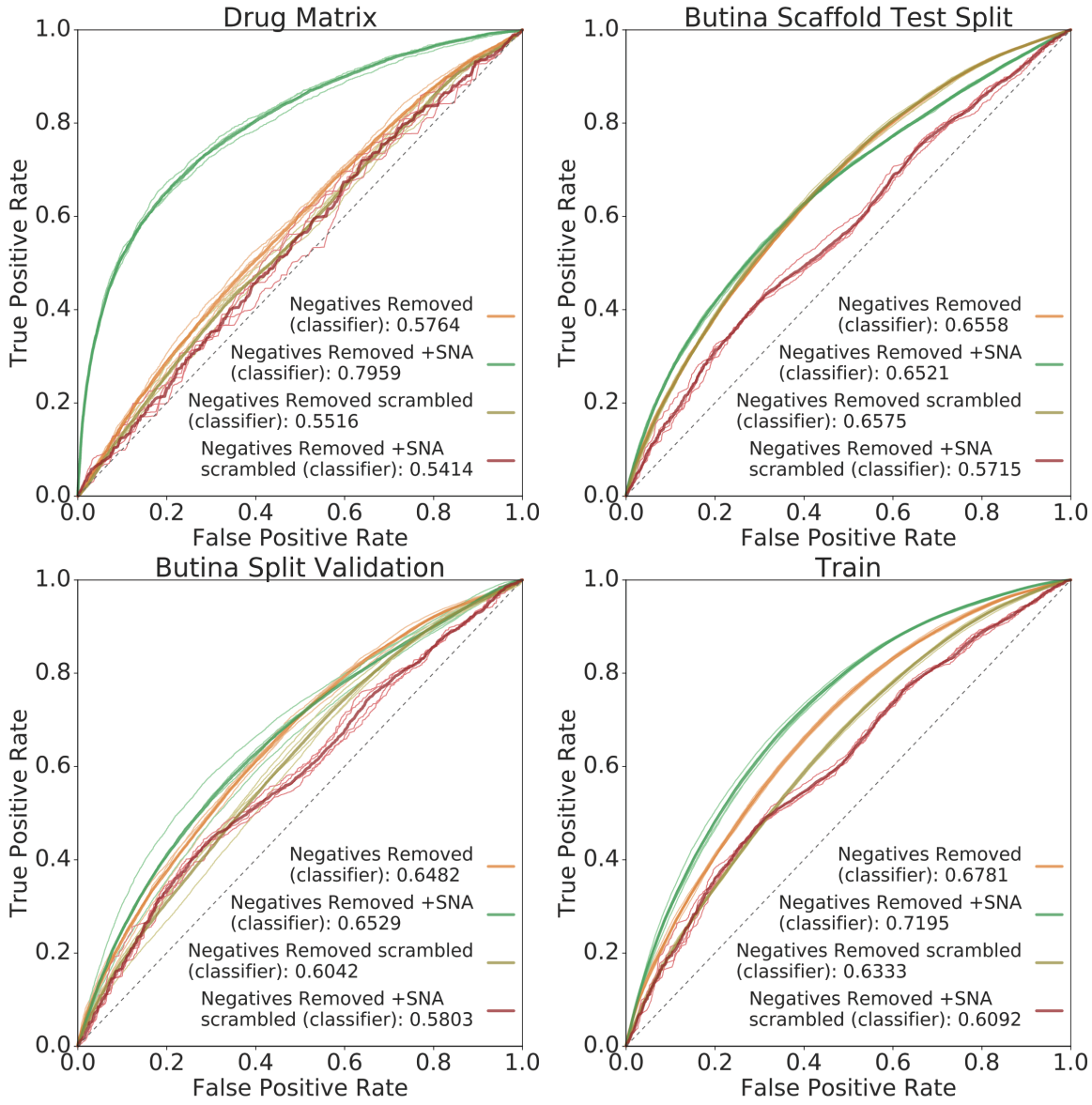

Figure A.34: Butina Split AUROC plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled classification DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC plotted with a thicker line.

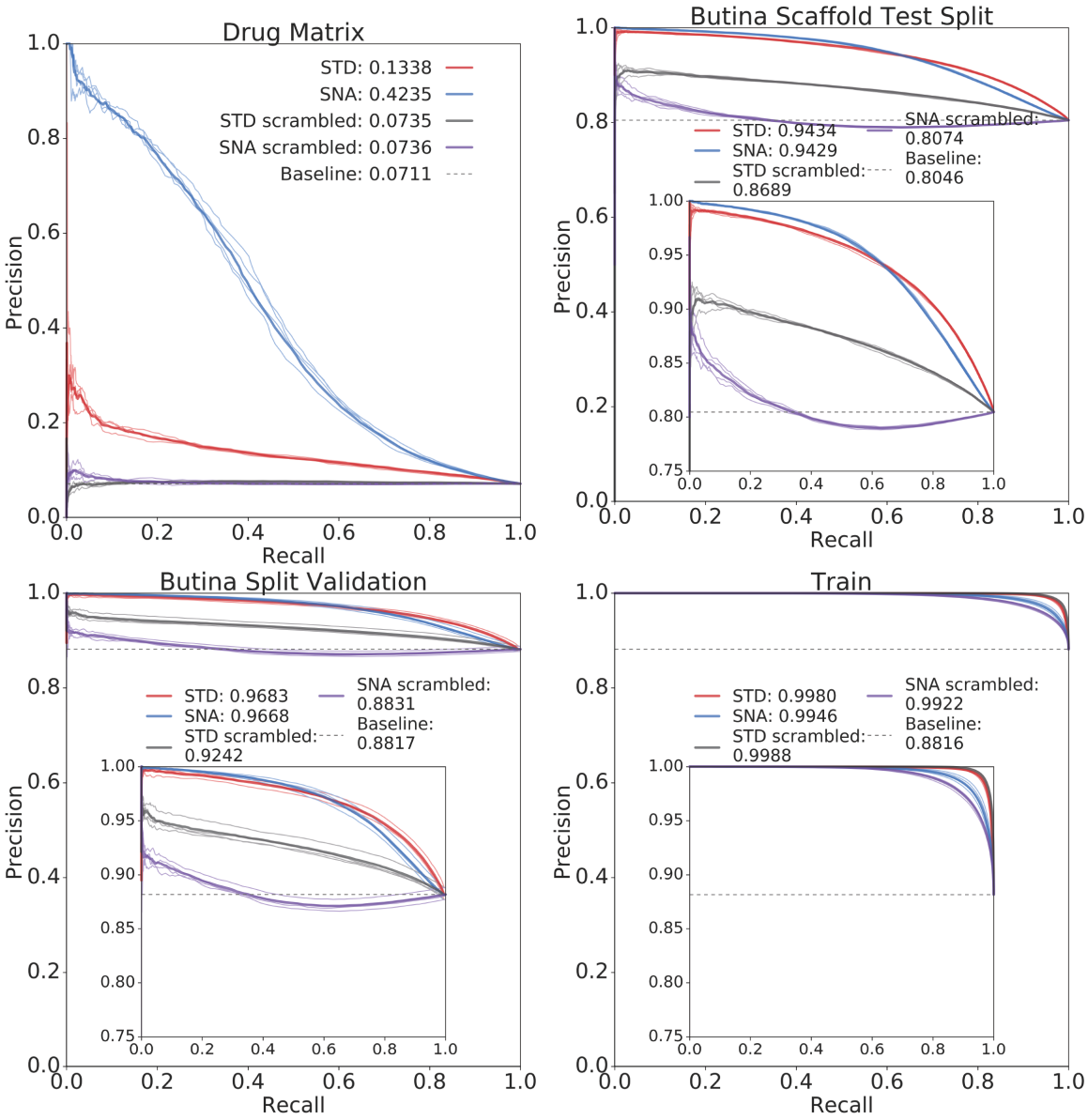

Figure A.35: Butina Split $AUPRC_r$ plots for SNA, STD, SNA scrambled, and STD scrambled regression DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean $AUPRC_r$ plotted with a thicker line.

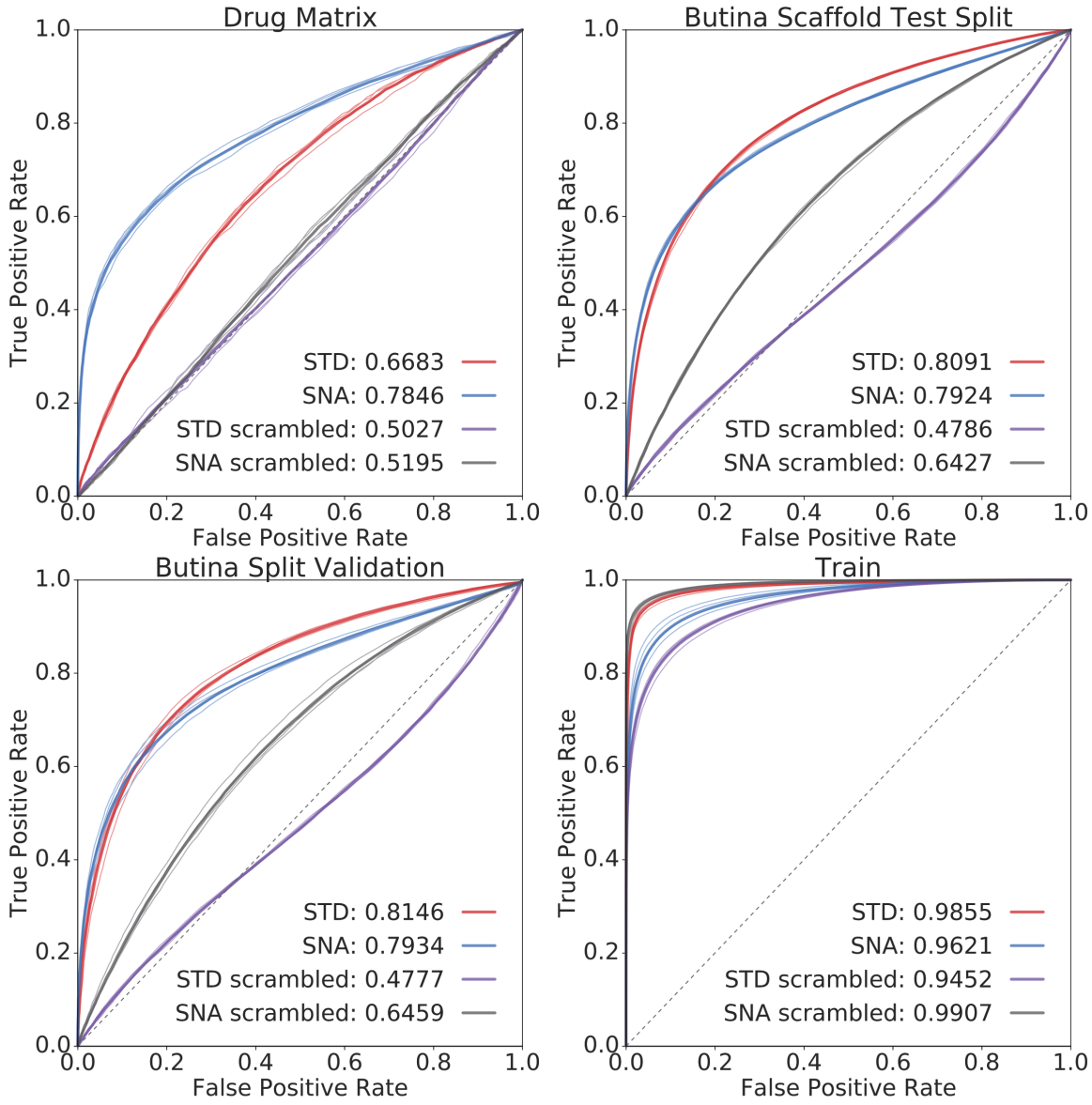

Figure A.36: Butina Split AUROC$_r$plots for SNA, STD, SNA scrambled, and STD scrambled regression DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC$_r$ plotted with a thicker line.

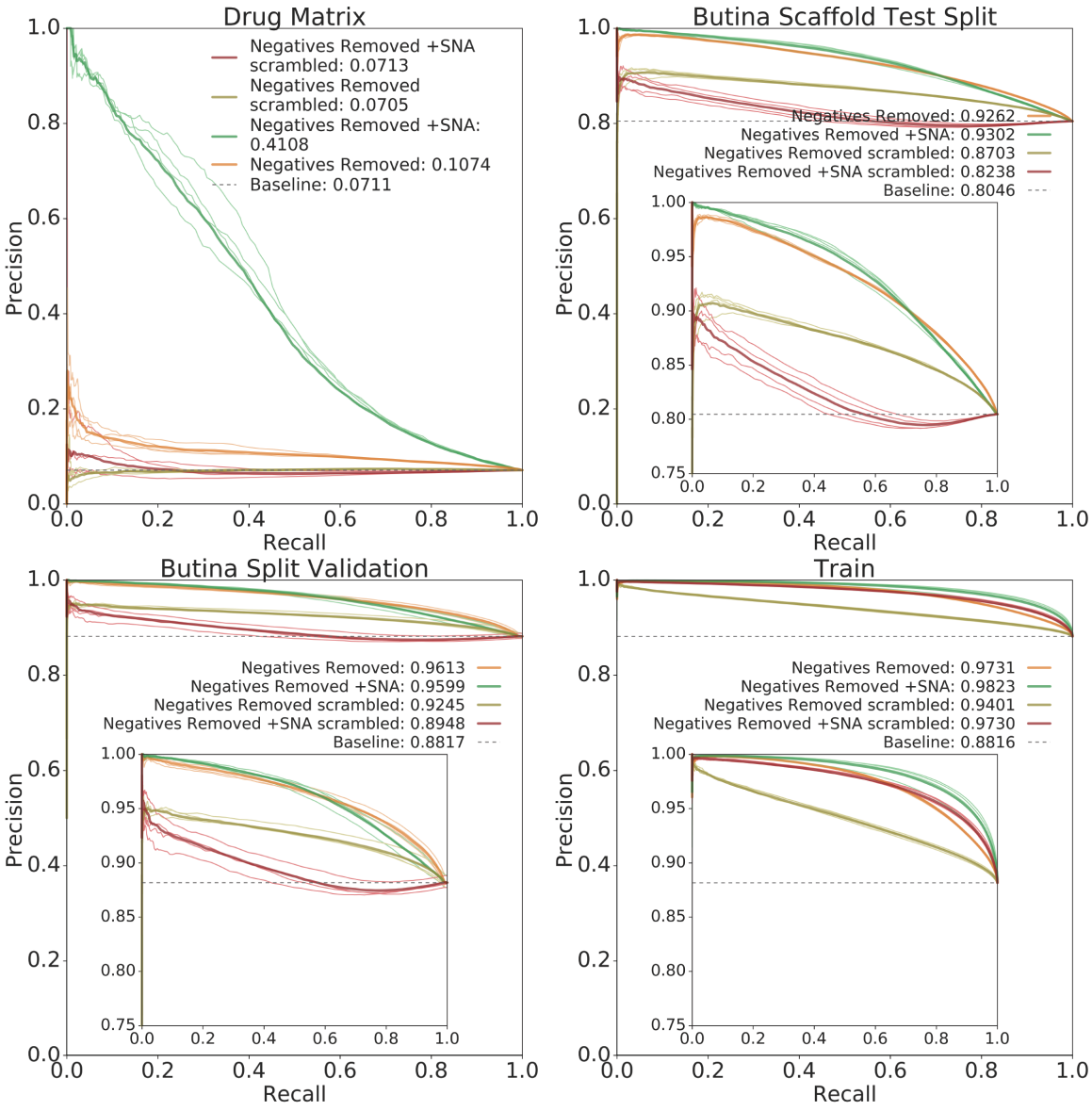

Figure A.37: Butina Split AUPRC$_r$ plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUPRC$_r$ plotted with a thicker line.

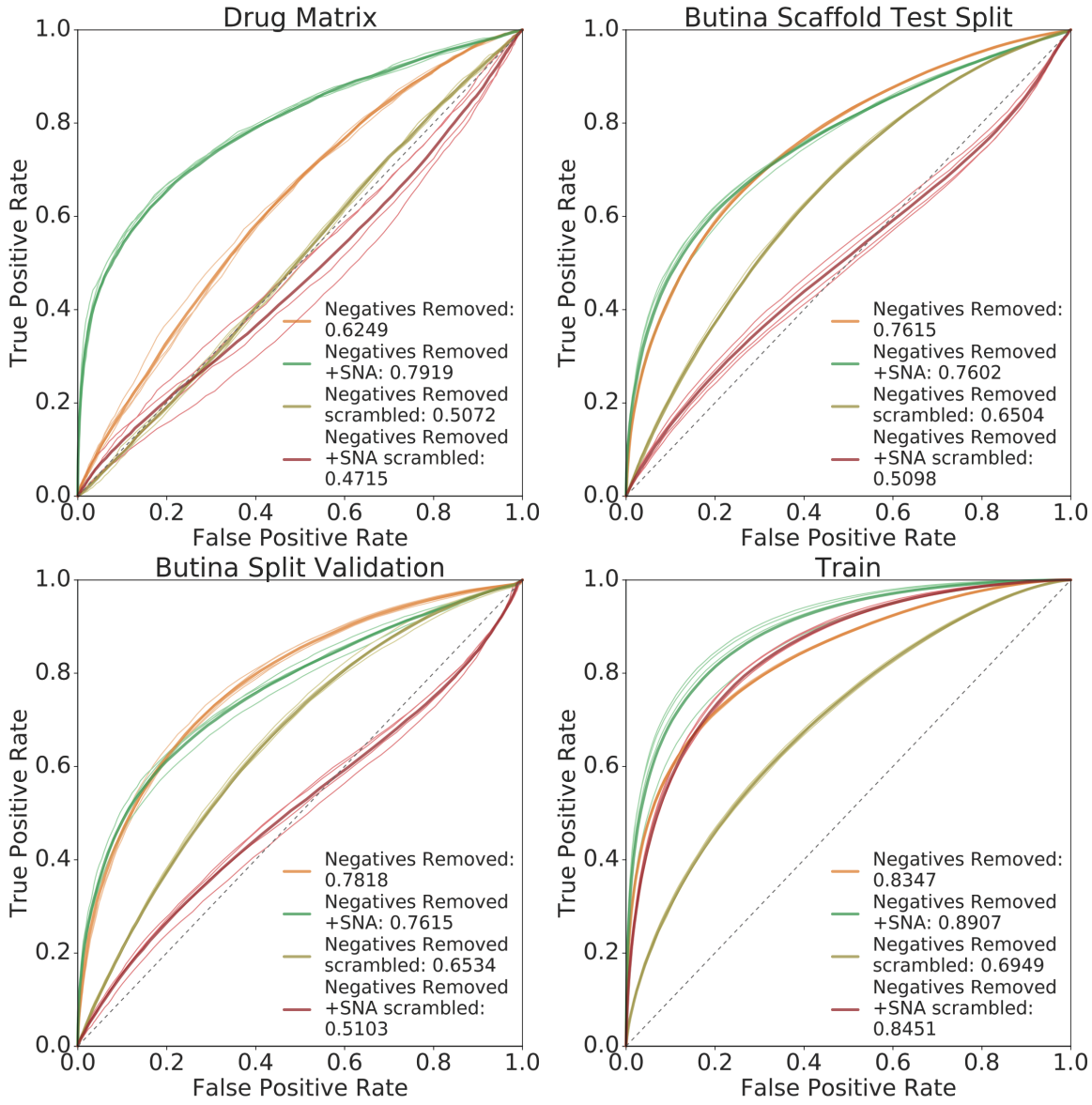

Figure A.38: Butina Split AUROC_r plots for Negatives Removed, Negatives Removed +SNA, Negatives Removed scrambled, and Negatives Removed +SNA scrambled DNNs for Drug Matrix (upper left), Butina Scaffold Test Split (upper right), Validation (lower left), and Train (lower right). Each fold is plotted individually, with the mean AUROC_r plotted with a thicker line.

References

1. Dalke, A. The chemfp project. *Journal of Cheminformatics* **11,** 76. ISSN: 1758-2946. doi:10.1186/s13321-019-0398-8 (Dec. 2019) (cit. on p. 123).

2. Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *Journal of Chemical Information and Computer Sciences* **35,** 59–67. doi:10.1021/ci00023a009 (1995) (cit. on p. 123).

3. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences* **39,** 747–750. doi:10.1021/ci9803381 (1999) (cit. on p. 123).

4. Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K. & Pande, V. MoleculeNet: a benchmark for molecular machine learning† †Electronic supplementary information (ESI) available. See DOI: 10.1039/c7sc02664a. *Chemical Science* **9,** 513–530 (2018) (cit. on p. 124).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Elena Caceres
714419733E5F4D7...
Author Signature

6/1/2021
Date