

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Machine Learning for Medical Image Analysis and Compound-Target Interactions

Permalink

<https://escholarship.org/uc/item/8c54b5m7>

Author

Gaskins, Garrett

Publication Date

2020

Peer reviewed|Thesis/dissertation

Machine Learning for Medical Image Analysis and Compound-Target Interactions

by
Garrett Gaskins

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:

DocuSigned by:

Michael J Keiser

Michael J Keiser

4DF1BD06D670465...

Chair

DocuSigned by:

Jason Gestwicki

Jason Gestwicki

DocuSigned by:

Sourav Bandyopadhyay

Sourav Bandyopadhyay

F1E602E27884493...

Committee Members

Copyright 2020

by

Garrett Gaskins

ACKNOWLEDGEMENTS

The graduate process is quite an experience. There is a subtle difficulty in describing it, as a peculiar quality of its existence is defined by the uniqueness of it all. Everyone I've known to be a part of or to have gone through graduate school has described a truly different journey. Perhaps this is integral to what makes the whole thing work. At the least, I can assert that the process, in general, is marked by growth. Throughout my time here, I have witnessed growth in my peers, in my mentors, and, through the help of others, friends, and family, in myself. Cliché perhaps, but nonetheless true. Here, I'd like to give thanks to those who helped me to grow during my graduate experience.

Firstly, I would like to thank my mentor, Michael Keiser, without whom I would not have had this opportunity. I joined Mike's lab as the first full-time student, and it has been a pleasure working with him since the beginning. Mike has always been available to discuss how best to tackle a problem, scientific or otherwise. For his understanding, patience, and active contribution to my development as a speaker and scientist, I am grateful. Watching the growth of Mike, as a mentor, leader, and PI, as well as the growth of the lab in general, has been a pleasure.

In addition to Mike, I'd like to thank the members of the lab for contributing to my scientific development and for tolerating my constant assault on olfaction. A special thanks to Elena Caceras, for her puns and friendship, Leo Gendelev for keeping me grounded in reality, Nick Mew for his help and strange choice in candies, Laura Gunzales for existential conversations and problem-solving skills, and Kangway Chaung for setting a new precedent of scientific rigor and mentorship in the lab, and schooling me in basketball.

A shout out to the members of the Gestwicki and Kampmann labs for always being helpful in and out of the wetlab space, and for providing a different window into how science can and should be conducted.

I'd like to thank my committee members Jason Gestwicki and Sourav Bandyopadhyay for their insightful input, their guidance in framing the right questions to ask for my project, as well as their time and expertise.

I cannot do justice to how much support my friends and family have given me over the years, but I can at least mention some of my favorite people that have helped me to get where I am. A special thanks to Clint and Nima for keeping me sane. Brian O'Donovan (and Danielle) for keeping me alive, and also for almost getting me killed! Max Messinio for cracking me up, Kojo, Lucas, and Kenny for keeping me occupied in and out of the gym. And Matt McCarroll for talking music and science.

My brothers Gavan, Gage, and Griffin for all their trash talking and support. My father for giving me his humor and wit, and my mother for all of her love and consideration.

Above all else, I thank Stephanie See for being an endless source of love, patience, and support, and for showing me a slice of life I never would have experienced otherwise. I couldn't have done this without you, and I so grateful to be able to experience new adventures with you and the dinosaur.

The text of Chapter 2 contains reprinted or adapted material from the previously published manuscript:

Irwin JJ, Gaskins G, Sterling T, Mysinger MM, Keiser MJ. Predicted Biological Activity of Purchasable Chemical Space. *J Chem Inf Model*. 2018 Jan 22;58(1):148-164.

John Irwin conceived of, funded, drafted, and performed the initial predictions for the paper. Teague Sterling helped build the ZINC database used to extract molecules. Michael Keiser provided insight, guidance, funding, and editing for the manuscript.

Machine Learning for Medical Image Analysis and Compound-Target Interactions

Garrett Gaskins

ABSTRACT

Though the title of my thesis infers a unifying theme via the application of machine learning, the two projects that form the bulk of my graduate degree are frankly more disparate than they are similar. Both endeavors provide novel methods to a field where ground truth is obscure and/or limited, and both apply machine learning techniques in their methodologies. Those similarities notwithstanding, the scientific domains, technical applications, experimental designs, and overall goals remain independent. While having a thesis comprised from two independent parts may not be conventional, this is, as they say, not a bug but a feature. Working within (and occasionally across) two research domains has helped me to acquire a diverse skillset and has provided me with a better and broader understanding of machine learning practices for scientific research.

As this thesis is composed of two linked, but distinct projects, the abstract (and chapters) is divided in two. The first section details work related to large-scale predictions of purchasable chemical space, and the second summarizes a novel method for automating diagnosis of melanocytic atypia in human histopathological samples.

I. Large-Scale Predictions for Purchasable Chemical Space

There are now over 400 million compounds one can easily purchase from the ZINC database (zinc.docking.org). About 350 million (85%) of these compounds are affordable enough for the average

academic lab to conduct a ligand discovery project. However, the molecular targets (proteins) that these purchasable compounds bind and modulate—if any—are rarely known. Fewer than 1 million compounds (<0.25%) have been reported active in a target-specific assay according to public databases such as ChEMBL. In the absence of target activity information, the process of selecting compounds for general purpose screening will often be target-naïve.

To facilitate access to new chemistry for biology, my collaborator John Irwin and I generated predictions for all purchasable compounds in ZINC at the time. I explored methods for optimizing predictive performance of compound-target associations using ChEMBL’s bioactivity dataset (version 21) as a benchmark. Comparisons on cross-validation sets of the bioactivity dataset against several methods such as multinomial naïve-Bayesian classifiers revealed that the combination of the Similarity Ensemble Approach (SEA) with the maximum Tanimoto similarity to the nearest bioactive yielded the best performance. I verified the utility for several of these predictions, quantified target prediction biases inherent to the dataset, and provided thresholding suggestions to the user for controlling sensitivity and specificity of the predictions, as well as novelty of target-associations allowed.

II. Automating Diagnosis of Melanocytic Atypia: A Precursor to Melanoma in Situ

Melanocytic atypia, a biological precursor to Melanoma, is histopathologically challenging. Pathologist interobserver agreement for melanocytic atypia in standard (H&E) histology images is low, ranging from 33-68%¹³⁶⁻¹³⁷, with melanoma in situ (MIS) in particular contributing to diagnostic discordance. A lack of agreement among experts presents a challenge to any supervised learning task, where the utility of a learned function depends on the accuracy and reliability of labels used.

To circumvent the issue of discordance in labeling, I paired H&E histology images with contiguously cut tissue sections, immunohistochemically (IHC) stained for melanocytes. I developed a

deep-learning pipeline for automating diagnosis of melanocytic atypia using a custom dataset of paired whole slide images (WSIs) and trained convolutional neural networks to identify the presence of melanocytes in H&E sections, using information solely from paired tissue samples. Networks achieve strong performance on holdout patient datasets. For each network trained, I generated full-scale (20X magnification) high-resolution (pixel-wise) prediction heatmaps on holdout tissue sections (H&E), for pathological interpretation, and applied saliency mapping to show what networks attend to in H&E images. This pipeline aims to provide assistance to the clinical pathologist to reach better consensus regarding new MIS diagnoses in cutaneous biopsies.

Table of Contents

Chapter 1: Predicted Biological Activity of Purchasable Chemical Space.....1

1.1 Abstract	2
1.2 Introduction	2
1.2.1 Purchasable Chemical Space	2
1.2.2 Scarcity of Binding Information in Purchasable Chemical Space	3
1.2.3 Predicting Compound Bioactivity.....	3
1.2.4 Utility of Predictions	4
1.3 Results	5
1.3.1 Prediction Database.....	5
1.3.2 Sensitivity and Specificity of Predictions	5
1.3.4 Controlling Novelty of Predictions	6
1.3.5 Accessing Predictions by Target of Interest.....	7
1.3.6 Access by Gene Groupings	8
1.3.7 Benchmarks.....	9
1.3.8 Genes Lacking Commercially Available Ligands	10
1.3.9 Dark Chemical Matter.....	11
1.4 Use Cases	11
1.4.1 Use Case One	12
1.4.2 Use Case Two	12
1.4.3 Use Case Three	13

1.4.4 Use Case Four	14
1.4.5 Obtaining 3D Models.....	14
1.5 Discussion.....	15
1.5.1 Summary of Results.....	15
1.5.2 Summary of Interface.....	16
1.5.3 Prediction Relevance and Target Bias	17
1.5.4 New Starting Points	18
1.5.5 Liabilities and Limitations	18
1.5.6 Conclusion	19
1.6 Methods.....	19
1.6.1 Library Preparation	19
1.6.2 SEA Reference Library Construction	20
1.6.3 Database Loading.....	20
1.6.4 ChEMBL Cross-Validation.....	21
1.6.5 Interface	22
1.6.6 Caveats to 3D Models.....	22
1.7 Acknowledgements.....	22
1.8 Figures.....	24
1.9 Tables.....	35

Chapter 2: Automating Diagnosis of Melanocytic Atypia: A Precursor to

Melanoma in Situ	42
2.1 Abstract	43
2.2 Introduction	44
2.2.1 Deep Learning in Medical Image Analysis.....	44
2.2.2 Melanocytic Atypia is Histopathologically Challenging	44
2.2.3 Circumventing the Necessity for Pathologist Manual Labeling	45
2.3 Methods	46
2.3.1 Ethics Approval	46
2.3.2 Sample Cohort	46
2.3.3 Sample Preparation	47
2.3.4 Dataset Split	47
2.3.5 Tissue Extraction and Filtering	48
2.3.6 Stain Threshold Calibration	48
2.3.7 Image Alignment and Tiling	49
2.3.8 Denoising IHC Sections.....	50
2.3.9 Performance Assessment	51
2.3.10 Prediction Heatmaps	51
2.3.11 Saliency Mapping	52
2.4 Results	52

2.4.1 Stain Calibration	52
2.4.2 Melan-A Model.....	53
2.4.3 SOX10 Model	54
2.5 Discussion.....	55
2.6 Figures	56
2.7 Tables.....	64
References	66

List of Figures

Figure 1.1 - Comparative performance of SEA, SEA+TC, and a multinomial naive-Bayesian classifier (NBC) on ChEMBL cross-validation sets.....	24
Figure 1.2 - Predictions supported by evidence.	25
Figure 1.3 - Tools to display predictions for a gene and filter and sort them by MaxTc and pSEA.....	27
Figure 1.4 - Predictions available for 2629 genes.....	28
Figure 1.5 - Prediction counts and purchasable compounds.....	31
Supplementary Figure A.1.1 - Performance metrics for SEA+TC on ChEMBL cross-validation sets filtered for >5 ligand annotations per target.....	32
Supplementary Figure A.1.2 - Performance metrics for SEA+TC on ChEMBL cross-validation sets filtered for >50 ligand associations per target.	33
Supplementary Figure A.1.3 – Prediction bias.	34
Figure 2.1 - Image processing pipeline.....	56
Figure 2.2 - Stain threshold calibration for cross-stain performance metrics.	57
Figure 2.3 – Melan-A Melanocyte-detector performance.	58
Figure 2.4 - Truth Table Prediction Examples.	59
Supplementary Figure A.2.1 – Automated Tissue Extraction Toolkit.	60
Supplementary Figure A.2.2 – SOX10 Melanocyte-detector performance.....	61
Supplementary Figure A.2.3 - Truth table predictions from TCGA holdout dataset.....	62
Supplementary Figure A.2.4 – Saliency Mapping.	63

List of Tables

Table 1.1 - Drugs with No Binding Data in ChEMBL, Predicted by SEA or MaxTC, Corroborated by the Literature	35
Table 1.2 - Selected Plausible Predictions of Purchasable Compounds for Genes with No Purchasable Ligands in ChEMBL	38
Table 1.3 - Compounds with No Predictions: “Chemical Dark Matter”	40
Table 2.1 – SOX10 Whole Slide Image Dataset Overview	64
Table 2.2 – Melan-A Whole Slide Image Dataset Overview	65

Chapter 1: Predicted Biological Activity of Purchasable Chemical Space

Irwin JJ*⁺, Gaskins G*, Sterling T, Mysinger MM, Keiser MJ⁺

University of California San Francisco, Institute of Neurodegenerative Diseases, 675 Nelson Rising Lane,
San Francisco, California 94143, United States

* Co-First authors

⁺ Corresponding authors

1.1 Abstract

Whereas 400 million distinct compounds are now purchasable within the span of a few weeks, the biological activities of most are unknown. To facilitate access to new chemistry for biology, we have combined the Similarity Ensemble Approach (SEA) with the maximum Tanimoto similarity to the nearest bioactive to predict activity for every commercially available molecule in ZINC. This method, which we label SEA+TC, outperforms both SEA and a naïve-Bayesian classifier via predictive performance on a 5-fold cross-validation of ChEMBL's bioactivity data set (version 21). Using this method, predictions for over 40% of compounds (>160 million) have either high significance ($p_{SEA} \geq 40$), high similarity ($ECFP4MaxTc \geq 0.4$), or both, for one or more of 1382 targets well described by ligands in the literature. Using a further 1347 less-well-described targets, we predict activities for an additional 11 million compounds. To gauge whether these predictions are sensible, we investigate 75 predictions for 50 drugs lacking a binding affinity annotation in ChEMBL. The 535 million predictions for over 171 million compounds at 2629 targets are linked to purchasing information and evidence to support each prediction and are freely available via <https://zinc15.docking.org> and <https://files.docking.org>.

1.2 Introduction

1.2.1 Purchasable Chemical Space

The purchasable chemical space has roughly doubled every two and a half years since 1990, owing to steady progress in efficient parallel synthesis¹⁻⁸ and the synthesis of new building blocks. There are now over 400 million compounds one can easily purchase using ZINC,⁹ which covers 204 commercial catalogs from 145 companies. Each catalog is categorized by ease of purchase, and each compound in turn inherits a purchasability level from its catalog membership. The growth in catalog size is impressive, particularly among the make-on-demand catalogs. Purchasable compounds in the favored lead-like¹⁰ and fragment-like¹¹ areas have grown from 3 million and a half million in 2007 to 124 million and 9.2 million

today, respectively. Many vendors have incorporated the lessons of lead- and fragment-likeness in library design,⁴⁷ often filtering for PAINS.⁴⁸ About 340 million (85%) of these compounds are affordable enough for the average academic lab to conduct a ligand discovery project, retaining a price point around \$100 per sample or less. A further 60 million compounds are available at higher building-block prices, often \$400 USD or more and are included here for completeness. We find that synthesis plus delivery of make-on-demand screening compounds often takes little more than a month or so, just twice the time to source many in-stock compounds.

1.2.2 Scarcity of Binding Information in Purchasable Chemical Space

The molecular targets (proteins) that these purchasable compounds bind and modulate—if any—are rarely known. Fewer than 1 million compounds—less than 0.25%—have been reported active in a target-specific assay according to public databases such as ChEMBL¹² or other annotated collections indexed by ZINC.¹³ Investigators searching for testable ligands might not consider the remaining readily available compounds, as they are not annotated for targets and the sheer number of options can be daunting. In the absence of target activity information, the process of selecting compounds for general purpose screening will often be target-naïve, relying on chemical or physical-property diversity to sample chemical and property space, respectively.¹⁴ If information on target bias—the likelihood that a compound is more disposed to bind to a particular target or class of targets—were readily available, libraries more likely to cover biological targets of interest could be designed.

1.2.3 Predicting Compound Bioactivity

Systematically assaying every commercially available compound against every target is experimentally impractical, so prioritizing compounds through computational predictions is a pragmatic alternative. There are many methods for predicting biological activities by chemical similarity,^{15–36} here, we use two. The Similarity Ensemble Approach (SEA)^{37,38} predicts biological targets of a compound

based on its resemblance to ligands annotated in a reference database, such as ChEMBL.¹² SEA relates proteins by their pharmacology by aggregating chemical similarity among entire sets of ligands. By leveraging extreme value statistics, SEA filters out unreliable signals and normalizes the aggregate results against a random chemical background to predict the significance of pharmacological similarity. SEA has successfully predicted targets of marketed drugs,^{37–39} toxicity targets,⁴⁰ and mechanism of action targets for hits in zebrafish⁴¹ and *C. elegans*⁴² phenotypic screens. We also use the maximum Tanimoto coefficient⁴³ at 0.40⁴⁴ or better based on ECFP4 fingerprints⁴⁵ to inform predictions. Neither method generates models incorporating discrete chemotypes as do Naïve Bayes classifiers, for instance, but instead consider the molecule holistically. This is advantageous because the method can suggest molecules that do not conform to what has been highly weighted by precedent. Other methods such as Naïve Bayes⁴⁶ can explicitly weight for chemical substructures that are potentially important to bioactivity (“warheads”), and thus a future version might use such an approach to complement this work.

1.2.4 Utility of Predictions

To be useful for research, predictions should be accessible, searchable, and downloadable. An interface should allow access to predictions for each compound, as well as for each target, vendor, and gene. A mechanism to select more novel or more conservative predictions would cater to a wide range of requirements. And libraries should be downloadable in 2D formats for chemoinformatics as well as in popular 3D formats for docking screens.

The prospective user of such a resource expects some way to evaluate the predictions. As one proxy to assess this data set, we performed a retrospective 5-fold cross-validation on the ChEMBL bioactivity data set for our method as compared to SEA and a naïve-Bayesian classifier, at a variety of threshold parameters (**Figure 1.1; Supplementary Figures A.1.1 and A.1.2**). Second, in assessing performance, we encountered the observation that whereas the canonical targets of all but a few drugs are known,⁴⁷ hundreds of established drugs and investigational compounds nonetheless lack their

respective target annotations in ChEMBL. We turned this deficit to our advantage, by testing the method's prediction of targets for several such drugs, corroborating our predictions with the literature when available. Finally, as these predictions are based on protein–ligand annotations derived from ChEMBL, we expect that this method will be silent about chemotypes and targets not contained in this approximation of the public pharmacopeia.

1.3 Results

1.3.1 Prediction Database

The ZINC database contains 400 million commercially available organic molecules with molecular weight between 50 and 1000 Da, sourced from 204 commercial catalogs published by 145 companies. We have created a database of predicted biological activities for the 171 million compounds that had predictions and have made it freely accessible via ZINC (<https://zinc15.docking.org>) and our file server (<https://files.docking.org>). All predictions were computed using a combination of the Similarity Ensemble Approach (SEA)³⁷ and Tanimoto similarity calculations based on compound annotations derived from ChEMBL Version 21¹² (see Methods). We refer to this combinatorial approach as SEA+TC throughout the text.

1.3.2 Sensitivity and Specificity of Predictions

To enhance this resource's applicability to a broad audience, we sought to increase the specificity of predictions by using more stringent criteria for what constitutes an annotated ligand. In prior work we had used a 10 μ M affinity cutoff, but at this scale, we encountered flawed predictions that appeared to arise from similarity to weak binders, possible PAINS, or promiscuous aggregator compounds. Based on our experience with these encounters, we changed the baseline affinity threshold to 1 μ M and further

required activities of at least 100 nM for compounds containing PAINS patterns or being Tc 0.70 to any compound observed to aggregate.⁴⁸⁻⁵⁰

We adopted a statistical significance threshold of negative log SEA p-value⁵⁴ ($\text{pSEA} \geq 40$) and a MaxTc cutoff ≥ 0.40 guided by the work on belief theory from the Abbvie group.³⁴ MaxTc is complementary to pSEA as it provides a single-nearest-neighbor-molecule view of similarity, compared to SEA's global view arising from the ensemble of annotated ligands. To quantify how this bivariate threshold improves predictive capability, we evaluated the performance of SEA, SEA+TC, and a Naïve-Bayesian classifier (NBC) via 5-fold cross-validation of ChEMBL's bioactivity data set (version 21; **Figure 1.1**). SEA+TC's ability to correctly predict compound-target interactions as either positive (does bind) or negative (does not bind) outperformed both SEA and the NBC, as measured by the area under the receiver operating characteristic (AUROC) curve, (AUROC = 0.995, **Figure 1.1A**). Further, when predicting a compound-target interaction as positive, SEA+TC was correct in its prediction more often than SEA or the NBC, as indicated by its area under the precision-recall (AUPRC) curve (AUPRC = 0.684, **Figure 1.1B**). In performing this analysis, we additionally identified a more stringent bivariate threshold, which some users may wish to adopt. At a threshold of MaxTc ≥ 0.80 with $\text{pSEA} \geq 80$, the retrospective analyses achieve higher precision than the baseline threshold (**Figure 1.1A and B**, blue circle) at acceptable recall (pink circle). Users of the ZINC interface may choose thresholds to suit their needs.

1.3.4 Controlling Novelty of Predictions

In addition to controlling the sensitivity and specificity of predictions, the significance threshold (i.e., pSEA and MaxTc values)¹⁷ also influences the novelty of the predictions. Novel compounds can be desirable because they likely have unrelated off-target effects, which can help establish the signaling and toxicity role of a receptor, as well as selectively activate downstream signaling, which is important for many receptors such as GPCRs.³⁸ Accordingly, we designed the ZINC interface to help users rapidly

identify predictions with their desired precision. The user can control the MaxTc and pSEA limits, and each prediction can be compared with the most similar annotated actives (**Figure 1.2**) allowing side-by-side comparison. Each SEA prediction is accompanied by a pSEA to the set of actives and MaxTc to the nearest active. Clicking on the MaxTc value in the interface performs a real-time search for the most similar ligands annotated at 10 μ M or better for that target.

1.3.5 Accessing Predictions by Target of Interest

To find predictions for a given target using ZINC15 (zinc15.docking.org), the user may select Genes from the Biological dropdown menu to browse a listing of all genes and predictions (**Figure 1.3A**). In this work, we use genes and their identifiers as convenient shorthand for their protein products—or molecular targets. To find a specific gene, the user may type part of the gene name in the top right search bar, here SLC6, and click the blue search button on the top right. To display predictions for this gene, the user clicks on the link in the predictions column, here for SLC6A1 (**Figure 1.3B**). The user may for example use the subset selector to specify strong predictions (which we chose to mean pSEA = 80) and purchasability (**Figure 1.3C**). Some advanced features are currently only accessible by hand-editing the URL. Here, the user adds `table.html?sort=-maxtc` and `&maxtc-between=40+45` to display the information in a tabular format, to sort by decreasing MaxTc, and to select only predictions between MaxTc of 40 and 45, respectively (**Figure 1.3D**). We plan to make these API-level features available via a point and click interface soon. Documentation is available via the help pages <https://zinc15.docking.org/genes/help> and <https://zinc15.docking.org/predictions/help>.

Predictions are available for 2629 genes⁵¹ (**Figure 1.4**). The number of predictions per gene varies substantially, reflecting both the diversity of annotated ligands for the target as well as how well these chemotypes are represented in current vendor catalogs. For example, natural products and their analogs are often difficult to access synthetically and are therefore generally sparsely represented. At the high end of predictions per gene, the eukaryotic GPCRs D₄ dopamine receptor (DRD4), C–C chemokine

receptor type 3 (CCR3), and the voltage gated ion channels KCNK3 and KCNK9 each have over 4.8 million purchasable predicted ligands. The number of strong predictions ($pSEA \geq 80$) varies from over 500 000 for KCNK3 to as few as 9181 for DRD4. Filtering at $MaxTc \geq 0.60$ instead, corresponding to a precision exceeding 0.334 using ECFP4 fingerprints,⁴⁴ the predictions for these four genes varied from as many as 25,728 for DRD4 to as few as 8912 for KCNK9. At the other extreme of predictions per gene, fungal laccase-2 precursor (LCC2), human C-C chemokine receptor type 6 (CCR6), voltage-gated sodium channel $Na_v1.9$ (SCN11A), and fruit fly DNA topoisomerase 2 (TOP2) each had fewer than 50 predicted commercially available ligands. The small number of predicted ligands can often be explained by a paucity of reference ligands; here, SCN11A and CCR6 have only 1 ligand each at 10 μ M or better. Another reason for the lack of ligands is that the knowns are in an area of chemical space that is difficult to access synthetically, such as natural products for both SCN11A and CCR6.

1.3.6 Access by Gene Groupings

In addition to individual genes, predictions may also be accessed by groups of genes. This could be helpful if the investigator is looking for new aminergic GPCR ligands or ligands for voltage gated ion channels or simply wishes to ensure balanced coverage of major target classes in a library. The interface offers convenient ways to access gene groupings based on a protein classification scheme inherited from ChEMBL. There are 15 major target classes (**Figure 1.5A**) further organized into 42 target subclasses (**Figure 1.5B**). Thus, there are 67 million predictions for membrane proteins, of which 1 million are strong ($pSEA \geq 80$). Considered separately, there are 873,000 less chemically novel predictions having a Tanimoto coefficient ≥ 0.60 to an annotated active. At a higher level of granularity, there are 4.7 million predictions for epigenetic reader proteins, of which 2.4 million are strong predictions ($pSEA \geq 80$) and 38 000 are highly similar ($Tc \geq 0.60$). At the organism level (**Figure 1.5C**), 18 million ligands are predicted for specific bacterial targets, 1.0 million of which are stronger ($pSEA \geq 80$) and 92 000 of which are highly similar ($Tc \geq 0.60$). The user may select purchasable compounds based on this classification.

These compounds will resemble preceded bacterial protein inhibitors far more strongly than compounds selected at random. Ligands predicted for specific bacterial targets are available to browse interactively at <https://zinc15.docking.org/organisms/bacteria/genes/> or to download by gene at <https://files.docking.org/predictions/current/>. A plot of predictions per gene vs annotated ligands per gene shows a general trend toward more predicted ligands when more known ligands are available (see **Supplementary Figure A.1.3**).

1.3.7 Benchmarks

We predicted the targets of established drugs that nonetheless lack a protein binding affinity annotation in ChEMBL to benchmark our approach. We found hundreds of drugs, withdrawn drugs, and investigational compounds with target predictions that agreed with the literature. Fifty of these were selected and tabulated as illustration of our predictions (**Table 1.1**). Thus, the beta blocker bufetolol⁵² (ZINC101) is predicted to be a β 2 adrenergic receptor ligand with pSEA = 47 and MaxTc = 0.46 and to be a β 1 adrenergic receptor ligand with pSEA = 51 and MaxTc = 0.44. Aramidipine⁵³ (ZINC600803) is predicted for the calcium voltage-gated ion channel CACNA1C with pSEA = 121 and MaxTc = 0.75. Ancarolol (ZINC39) illustrates the discriminatory value of the SEA prediction, with pSEA = 59 and MaxTc = 0.43 for ADRB1: 255,656 purchasable ligands have higher MaxTc than ancarolol to this target while only 46,753 have a higher pSEA score.

Among the 535 million predictions of protein–ligand affinity we expect numerous false positives and false negatives. These errors stem from three major classes of problem. (1) Issues with target annotation: annotated ligands may not be representative for a gene, such as curcumin (ZINC100067274), which is annotated for 32 genes and is probably artifactual for many of them.⁵⁴ Annotated ligands may also be mis-annotations in ChEMBL, leading to false positives. For instance, nicotinamide (ChEMBL1140) is annotated for fatty-acid amide hydrolase 1 (FAAH), because it shares an abbreviation (NAM) with the actual ligand, N-arachidonylmaleimide.⁵⁵ (2) Errors with the SEA method: We use

ECFP4 fingerprints, which have little specificity for certain classes of molecules, such as peptides and sterols, which share many common features and thus are not well discriminated using this fingerprint. SEA also has high variance for small ligand sets and low sensitivity for large, diverse ligand sets. For instance, SEA fails to predict the well-known antihistamine drugs chlorcyclizine and propiomazine for histamine H1 receptor (HRH1), despite their having Tc values of 0.79 and 0.69, respectively, to the most similar HRH1 ligands. The pSEA values of 11 in each case have been diluted by the 9000 diverse ligands annotated to this target. A remedy might be to split targets with large number of ligands, perhaps by chemical clusters, mode of action, or binding site, if known. Note that Naïve Bayesian classifiers can be trained to correctly predict these activities, as can be seen on ChEMBL's ligand detail pages for these compounds. (3) No explicit model of promiscuity for SEA: We have made some progress here by stringent filtering of ligands we suspect are promiscuous (both PAINS and aggregator-like), but we fail to handle frequent hitters such as staurosporine (ZINC3814434, hits 365 targets in ChEMBL) and its ilk. Our current approach also performs poorly on sigma nonopioid intracellular receptor 1 (SIGMAR1) and cytochromes P450-3A4 (CYP3A4), because the ligands annotated to it are highly diverse. To remedy this problem for targets with many ligands, we could cluster by chemotype.

1.3.8 Genes Lacking Commercially Available Ligands

When a target has purchasable ligands, they can be used to rapidly probe its biological function without requiring synthetic chemistry expertise. Yet there are 69 targets with 20 or more annotated ligands in ChEMBL where none is readily purchasable (**Table 1.2**). To fill these holes in “target space”, we have identified purchasable compounds that are predicted to be active. In one example, voltage dependent calcium channel subunit alpha-2/delta-2 (CACNA2D2) has 26 ligands in ChEMBL, none of which is for sale, such as CHEMBL1801206 with a pKi of 7.7. The compound ZINC36664273, however, is sold by Specs as AO-476/43421055 and has a pSEA of 132 and a MaxTc of 0.72. Looking at these compounds side by side (**Table 1.2**) and without detailed experimental knowledge of this target, the

Specs compound may be reasonable to try against this target. If successful, such compounds could become a purchasable control for these targets.

1.3.9 Dark Chemical Matter

Intriguingly, 229 million purchasable compounds have no prediction at all by either $pSEA \geq 40$ or Tanimoto similarity $Tc \geq 0.40$. Some of these will have just missed our cutoffs, wherever the cutoffs may be drawn. A few will be known actives, or analogs of actives, that simply lack a direct binding annotation in ChEMBL. Still, these compounds are generally interesting because they do not much resemble any direct binding actives in ChEMBL. Should they be found to be active in an assay, they are more likely to have fewer off-targets, at least against well-studied targets, and are less likely to be encumbered by patents. A substantial body of literature explores the strengths and pitfalls of dark chemical matter.⁵⁶⁻⁵⁹ To illustrate what a user of this resource can expect to find in this underexploited yet commercially available space, we have highlighted ten compounds (**Table 1.3**). For each commercially available molecule, we show the nearest precedented bioactive from public sources available to ZINC, which may also include compounds not in ChEMBL. Dark chemical matter⁵⁶⁻⁵⁹ may be browsed online at zinc15.docking.org/substances/having/no-predictions and downloaded at scale by physical property tranches (<https://files.docking.org/dark-matter/current>), by vendor catalogs (e.g., for ChemBridge at <https://files.docking.org/catalogs/50/chbr/chbr.predict.txt.gz>) and by the genes they are predicted to bind (<https://files.docking.org/genes/<genesymbol>/<genesymbol>.predictions.txt.gz>).

1.4 Use Cases

A new research tool is now available within ZINC15 for public use. We demonstrate the use of these new tools in four use cases, which illustrate how to access predictions both interactively and via static downloads, below.

1.4.1 Use Case One

The user is interested in a well-studied target such as the serotonin 2A receptor (HTR2A) and seeks compounds to purchase that are likely to work but have not been reported active in ChEMBL21. The user first checks how many ligands are annotated active at 10 μ M or better (5031, interactively at <https://zinc15.docking.org/genes/HTR2A/substances> or statically downloaded at <https://files.docking.org/genes/current/HTR2A/HTR2A.smi>). The user then queries how many commercially available ligands have SEA predictions at an exceptionally strong statistical significance, with $pSEA = 80$ (30,952 at <https://zinc15.docking.org/genes/HTR2A/predictions/subsets/strong+purchasable>). For instance, ZINC462039162 available from Enamine, catalog number Z1269906839, with a $pSEA = 82$ and $MaxTc = 0.63$ (<https://zinc15.docking.org/substances/ZINC000462039162/predictions/table.html>). Millions of other commercially available molecules can be obtained in a similar way. All predictions are downloaded immediately using <https://files.docking.org/genes/current/HTR2A/HTR2A.predictions.txt.gz>, from which compounds may be selected.

1.4.2 Use Case Two

The user wishes to obtain a screening library for projects involving several voltage-gated ions channels. The user wishes to find purchasable compounds that do not seem too similar, yet are more likely to be ligands than purely random compounds, i.e., having a high $MaxTc$ between 0.65 and 0.70, corresponding to an expected precision of 0.35–0.40. The library should be downloaded in 2D for chemoinformatics and 3D for docking. In ZINC, there are 14 849 already annotated ligands for any such channel in ChEMBL21 at 10 μ M or better (<https://zinc15.docking.org/subclasses/vgic/substances>). Of these, 1108 (7.5%) are purchasable and may be a good starting point for the library. A further 21 242 purchasable predicted ligands also are available, such as ZINC629100

(<https://zinc15.docking.org/substances/ZINC000000629100/predictions/table.html>), which is Tc 0.69 to the nearest annotated active ChEMBL1097858, active at pKi of 7.7. To obtain the first 1000 ZINC codes for these molecules, the user accesses:

<https://zinc15.docking.org/subclasses/vgic/predictions/subsets/purchasable.txt?maxtc-between=65+70&count=1000>. To download 3D models of these compounds, please see Obtaining 3D Models, below. A second approach to download predicted compounds for voltage gated ion channels would be to first obtain the names of all the genes:

<https://zinc15.docking.org/subclasses/vgic/genes.txt:name>. Then, the user would use this list to download the static predictions by gene. For example, for the sodium channel protein type 5 subunit alpha (SCN5A), the predictions are in <https://files.docking.org/genes/SCN5A/SCN5A.predictions.txt.gz>.

1.4.3 Use Case Three

The user would like to know all predictions for a particular vendor catalog. Vendors may be interested to know possible targets of their compounds for marketing purposes. Vendors may also wish to know which of their make-on-demand compounds might be prioritized for synthesis based on possible activity. Academic centers that screen vendor libraries may be interested in individual vendors because they have negotiated special pricing, or because the vendor makes plates available at a discount to facilitate the mechanics of screening. We have been precomputed searches to enable such investigations to save time. To access them, the user would complete the following steps:

1. Browse to <https://files.docking.org/catalogs> to select the catalog of interest.
2. Download the file of predictions. For instance, for ChemBridge, the code is chbr and the URL is <https://files.docking.org/catalogs/50/chbr/chbr.predict.txt.gz>. Each row contains the vendor code, ZINC ID, InChIKey, predicted gene, MaxTc, and pSEA: one molecule per row.

3. Break the downloaded files into subsets using Unix command-line tools to filter by MaxTc, pSEA, and predicted gene.

To download these in 3D for docking, please see Obtaining 3D Models, below.

1.4.4 Use Case Four

The user wishes to download dark chemical matter screening libraries in 2D or 3D formats. To do so, the user browses to <https://files.docking.org/dark-matter>. The compounds have been binned into tranches by physical property using our standard scheme (http://wiki.docking.org/index.php/Physical_property_space). The 2D files are available as compressed text files organized by purchasability. Each row contains one molecule with its SMILES, ZINC ID, physical property tranche, purchasability, and reactivity. The 3D files will likewise be prepared in future but are meanwhile available as described in Obtaining 3D Models, below.

1.4.5 Obtaining 3D Models

To download 3D models for a set of molecules in bulk for one of the above use cases, here is a general approach that will work for any arbitrary set of ZINC IDs:

1. Obtain the codes of the molecules to download using the previous use cases or otherwise and store the codes in `zinc-codes.txt`.
2. Select mol2, db, or db2 file formats. mol2 may be converted to other formats as required. The latter two are used by the UCSF DOCK 3.x programs only.
3. Download the script `getfiles.csh` from <https://files.docking.org/catalogs/getfiles.csh>.
4. Edit the file by hand following the instructions within.
5. Run the script, with the list of ZINC codes in the same directory. The 3D files will be downloaded.

Please note that 3D models are currently available for about 120 million of the 400 million compounds in ZINC. We are continually building and rebuilding them, prioritizing the popular lead-like and fragment-like areas best suited to docking. If a 3D model is not available, the molecule detail page contains a “Request Generation” button in the 3D representations section. If a 3D model does not exist, it is either because it fails to build or because it is still on our action list.

1.5 Discussion

1.5.1 Summary of Results

Four major results emerge in this work. First, using ZINC and ChEMBL, we predict molecular target activities for 171 million commercially available compounds at 2629 targets and store them in an accessible database. Second, we create an interface to search, access, and download the predictions (<https://zinc15.docking.org> and <https://files.docking.org>). Predictions can be accessed individually or downloaded in bulk, and are available in a range of formats ready for both docking and cheminformatics, or for purchase. To demonstrate the utility of these predictions, we perform a retrospective 5-fold cross-validation of the ChEMBL bioactivity data set. Further, we identify likely targets of drugs known in the literature where direct binding annotations are not available in ChEMBL. Finally, this new tool allows us to quantify predicted target biases of purchasable chemical space. Target bias predicted by this model is substantial—some genes are represented by millions of purchasable compounds, others have very few. Nearly 60% of purchasable compounds in ZINC have no prediction at all, allowing us to offer purchasable “dark chemical matter”. We take up each of these results in turn. We predict targets for over 40% of the 400 million compounds currently for sale in ZINC. The number is admittedly arbitrary, as we were obliged to choose pSEA and Tanimoto similarity cutoffs. Knowing that this approach would produce false positives and false negatives, we attempted to strike a useful balance,

and equip the user to apply further constraints. Many compounds with MaxTc as low as 0.40 to the nearest active may not bind the predicted target—previous work suggests 18% precision might be a good estimate⁴⁴ and this is consistent with the results we found in Figure 1 (blue circle). Likewise, those with a pSEA near our chosen threshold of pSEA = 40 may not be active against the predicted target. Should such chemically novel predictions be confirmed experimentally, they may represent new starting points for optimization and could lead to new biology. If the user wishes higher confidence hits, more stringent cutoffs in pSEA or MaxTc are easily applied. We refer the reader to the set of thresholds examined in our cross-validation of the ChEMBL bioactivity data set (**Supplementary Figure A.1.1**) for guidance in choosing pSEA and MaxTc values to optimize the desired output. For the highest rates of precision at an acceptable recall, we recommend threshold values at pSEA \geq 80 and MaxTc \geq 80 (**Figure 1.1**, pink circle), noting this may reduce the number of novel compound–target associations that pass the cutoff. For those wishing to buy a compound that works, the user might only consider the most similar compounds, having high Tc to a precedented bioactive. For those seeking chemical novelty against a target, where testing 10 or even 50 more novel compounds to find new chemical matter is acceptable, more novel compounds may be sought. Users of virtual screening methods such as docking may want particularly novel (low MaxTc) compounds, because their screening method makes an independent assessment of each prediction. Some will prefer to pursue the most novel—and potentially most interesting—the purchasable chemical dark matter, those compounds that do not seem similar to any of the annotated compounds used to make these predictions. Whatever the appetite for risk, investigators are empowered by these tools to select predictions that are right for their project.

1.5.2 Summary of Interface

Interfacing the prediction database through ZINC allows predictions to be searched, grouped, filtered, compared, and downloaded using the extensive ZINC machinery. Thus 3D models of predicted compounds may be accessed for molecular docking screens, while SMILES strings or molecular

properties may be downloaded for ligand-based methods. Predicted compounds for any of 2629 genes may be accessed and downloaded in any of eight formats. Results may be filtered by prediction statistics (pSEA, MaxTc), molecular properties (e.g., molecular weight, calculated logP, polar surface area, fraction sp³) and purchasability (in stock, make-on-demand, or by vendor). Both 2D and 3D results can be organized by gene (e.g., ADRB2, SRC), minor class (e.g., GPCR Class B, voltage-gated ion channel), major class (e.g., transcription factor or membrane protein), Kingdom (bacterial, eukaryotic, viral), vendor, and physical property tranche. Attributes of predictions may be downloaded in tabular form for analysis. A REST API, exemplified in this work, described previously⁹ and documented online,⁶⁰ allows automated queries and machine-readable results, so that this database may be incorporated into third-party software applications.

1.5.3 Prediction Relevance and Target Bias

We examined drugs and investigational compounds without an established molecular target annotation in ChEMBL to assess the relevance of the predictions. The 50 we highlighted exemplify typical results that can be expected using our approach for the millions of molecules that have never been assayed (**Table 1.1**). Whereas an exhaustive analysis is impractical, this result supports the view that our predictions are often consistent with experimentally observed binding.

A global picture of target bias in commercially available libraries emerges. Of the 535 million compound–target predictions, over 500 000 predictions on 400 000 compounds have a MaxTc better than 0.60 (ECFP4) to a ligand annotated for that target; a level of similarity that suggests 35% precision.⁴⁴ A further 1.6 million predictions on 1.4 million compounds with MaxTc between 50 and 59 are also strong candidates for experimental testing. Many of these two million compounds could have been predicted by pairwise Tanimoto similarity alone, without the help of SEA. The pSEA adds most value below MaxTc 0.50, where it provides a global similarity measure to the set of annotated ligands as a group instead of a single pairwise one. This becomes even more acute below MaxTc of 0.40, where we only retain

predictions with $pSEA \geq 40$ as the Tanimoto coefficient alone becomes too untrustworthy, with precision falling rapidly below 10%.

1.5.4 New Starting Points

Our analysis provides additional resources. We have predicted compounds for 69 targets⁶¹ for which none of the 20 or more actives is commercially available (**Table 1.2**). If confirmed experimentally, these genes could now be represented in screening panels of commercially available compounds, and these new ligands used as controls or perhaps even starting points for design. For each of 2629 genes, a range of commercially available compounds from high-confidence, having high MaxTc, to more-novel-yet-intriguing at lower MaxTc are now available. For the most studied targets, there is a deep bench of predictions running into the millions of compounds each. Massive biases for some targets, such as the dopamine D2 (DRD2) and beta-2 adrenergic (ADRB2) receptors for instance, echoes our earlier work⁶² that commercial libraries are heavily biased toward long-studied, important biological targets. Correspondingly, less-well-studied targets with few ligands often have sparse representation in commercial libraries, which can occur when the known actives are natural products or their derivatives. We have also assembled a database of “dark chemical matter”, 229 million purchasable compounds that received no target prediction and that generally do not resemble known bioactives, which is available from our website in 2D and 3D formats. If these compounds were active in a screen, they would likely represent new starting points for optimization.

1.5.5 Liabilities and Limitations

Our approach has other liabilities. Our cutoffs in MaxTc and pSEA inevitably exclude sensible predictions. Some classes of compounds such as sterols, peptides, and nucleotides suffer from higher misprediction rates, a subject of continuing research. pK_a and explicit charge are poorly treated in our current protocol based on stereochemistry-naïve ECFP4 fingerprints, making amide nitrogens and basic amines

too much alike, for instance, leading to some obviously wrong predictions. Massive turnover in the chemical marketplace means stored predictions may lag the appearance of new compounds in ZINC. ChEMBL contains artifacts and errors, which this approach can magnify. The SEA and MaxTc approaches quantify whole-molecule similarities and are thereby naïve of critical chemical moieties (often called warheads).

1.5.6 Conclusion

Notwithstanding these limitations, our database of predicted biological activities for purchasable chemical space is a pragmatic tool that should be useful to a broad audience. It affords both a retail view—buy this compound for this target—as well as a wholesale one—this target is well represented, and here are some compounds for it. Our predictions can be rapidly tested because the compounds are purchasable. We intend to continue to update the database as purchasable chemical space evolves and ChEMBL is enhanced. This database is provided in the hope that it will be useful, but you must use it at your own risk.

1.6 Methods

1.6.1 Library Preparation

We used ChEMBL21 compounds annotated for targets better than 10 μ M and grouped by Uniprot gene symbol across eukaryotes, as previously described in ZINC15.⁹ Thus in this scheme, DRD2_HUMAN, DRD2_RAT, and DRD2_MOUSE are all grouped into a single gene annotation DRD2, and predictions are made against the unified collection for the gene and not the individual orthologs. In situations where the target is composed of several gene products, as in some ion channels for instance, we used the ChEMBL name. When no gene has been formally assigned by Uniprot, we use the Uniprot accession code itself as the gene name, as in ZINC15.

1.6.2 SEA Reference Library Construction

We grouped ligands by affinity. We computed an affinity bin as the negative log of the molar affinity, which is variously expressed as K_i , IC_{50} , and EC_{50} among others in ChEMBL21 and which we refer to as pKi in this work for simplicity. Thus in this scheme, bin 6 contains all compounds with 1 μ M affinity or better. Lower affinity bins were inclusive of compounds from all higher affinity bins. We built three SEA libraries as follows. In the first library, we only proceed if there are at least five distinct compounds active against a single gene, we only accept activities of 1 μ M or better. We found 1382 such genes, which we defined as being well described by their ligands. In the second library, we only predict for those single gene targets that did not qualify for the first pass, accepting activities as weak as 10 μ M, and as few as one good ligand. We found 1347 of these less-well-described genes. The third library was an attempt to overcome a statistical weakness, which diluted the signal of genes having many diverse ligands. We clustered ligands to describe individual chemotypes of 302 genes having 300 ligands or more each. For each library we computed a statistical background for SEA based on the 410,624 annotated compounds. We computed the pSEA based on an extreme value distribution and the maximum Tanimoto similarity of the prediction to the annotated compounds (MaxTc). Throughout we suppressed from the libraries compounds with PAINS patterns or similarity to a precedented aggregator by 0.70 (ECFP4) having an affinity worse than 100 nM.⁴⁸ This was likely too conservative, but earlier, more permissive attempts at this library often suffered from excessive erroneous predictions, likely owing to these fraught compounds.

1.6.3 Database Loading

Predictions were loaded into ZINC. To minimize ligands whose charge differed sharply from precedent, we computed the mean and the standard deviation of the average microspecies charge using ChemAxon's CXCALC program for each gene. When loading each prediction, if the charge of a 3D representation at pH 7.4 (reference model) was available, we suppressed loading if the charge on the

molecule fell outside 1.5 standard deviations from the mean charge for ligands annotated to that gene. This remains an area of ongoing research. The result was to suppress predictions that we likely would have thrown out on inspection, in a scalable if incomplete and imperfect way.

1.6.4 ChEMBL Cross-Validation

We evaluated the predictive performance of SEA+TC using ChEMBL's bioactivity data set (version 21). Receiver operating characteristic curves were generated from independent 5-fold cross-validation runs for each method examined (SEA, SEA+TC, NBC). For SEA and NBC cross-validation sets, each point on the curve represents the average true-positive rate (TPR) and false-positive rate (FPR) from all 5 folds. TPRs and FPRs along the curve were determined by stepping a decision threshold across the range of possible SEA p-values (0.0–1.0), for all predicted compound-target interactions. To examine the sensitivity of these results to how well the target is described by ligands, we ran the analysis using targets with a minimum of 5 ligands and also with 50 ligands.

For SEA+TC cross-validation sets, TPRs and FPRs along the curve were determined by two separate decision thresholds; one for the SEA p-value and another for the maximum Tanimoto coefficient (MaxTc). As ROC curves evaluate a binary classifier using a single discrimination threshold, assessing performance by simultaneously stepping across both metrics was not ideal. To account for this, we generated ROC curves by stepping across all possible values of MaxTc, while holding the pSEA decision threshold constant (**Figure 1.1**). Predicted compound–target associations are therefore positive if their pSEA or MaxTc passes either of the respective cutoffs. A consequence of this bivariate thresholding is that the static pSEA threshold prevents the TPR and FPR from ever reaching zero. To highlight this, the distance between a fully stratified classifier (TPR = 0; FPR = 0) and the minimum point at which both decision thresholds begin to affect performance is shown in dashed lines (**Figure 1.1**). Performance metrics for a range of pSEA decision thresholds are shown in Supplementary Figure A.1.1, A and B.

Complementary curves stepping across pSEA while holding a separate MaxTc decision threshold constant are shown in Supplementary Figure A.1.1, C and D.

1.6.5 Interface

We added support for SEA predictions to the user interface on the Molecule Detail, Target Detail and Gene Detail pages of ZINC. The interface classifies each gene by one of 15 major target classes (e.g., membrane receptor, ion channel, transporter) and by one of 42 subclasses (e.g., Class A GPCR, voltage gated ion channel, etc) whose pages also allow access to the SEA predictions. The results are downloadable in eight formats: SMILES, mol2, SDF, pdbqt, json, xml, txt, and xls. The predictions may be accessed visually via a web browser or programmatically using an application program interface, both located at <https://zinc15.docking.org/predictions/home>. Static files are accessible via <https://files.docking.org/predictions>, <https://files.docking.org/genes>, <https://files.docking.org/catalogs>, and <https://files.docking.org/dark-matter>.

1.6.6 Caveats to 3D Models

Vendors often advertise stereochemically ambiguous molecular descriptions and thus the number of compounds and predictions strongly depends on how these are treated. Since ZINC is a 3D focused database, we are obliged to commit to a 3D representation. Where there is ambiguity, we enumerate up to a maximum of four possible stereoisomers (R/S and E/Z) and readily admit that this inflates the numbers in this work.

1.7 Acknowledgements

This work was supported by GM71896 (to B. K. Shoichet and J.J.I.), GM093456 (to M.J.K.), and the Paul G. Allen Family Foundation (to M.J.K.). We thank Greg Landrum, Novartis, and the RDKit

community for RDKit, ChemAxon (chemaxon.com) for licenses for JChem and Marvin, and OpenEye Scientific Software (eyesopen.com) for software licenses for OEChem. We thank SeaChange Pharmaceuticals (seachangepharma.com) for improvements to the SEAware software. We thank Dr. Matthew O'Meara for reading the manuscript. We are grateful the anonymous reviewers who made helpful suggestions that substantially improved the manuscript. While ZINC itself remains noncommercial, the contribution of a single company, Enamine Ltd (Kyiv Ukraine, <http://enamine.net>), offering over 80% of the compounds currently offered for sale, is acknowledged.

1.8 Figures

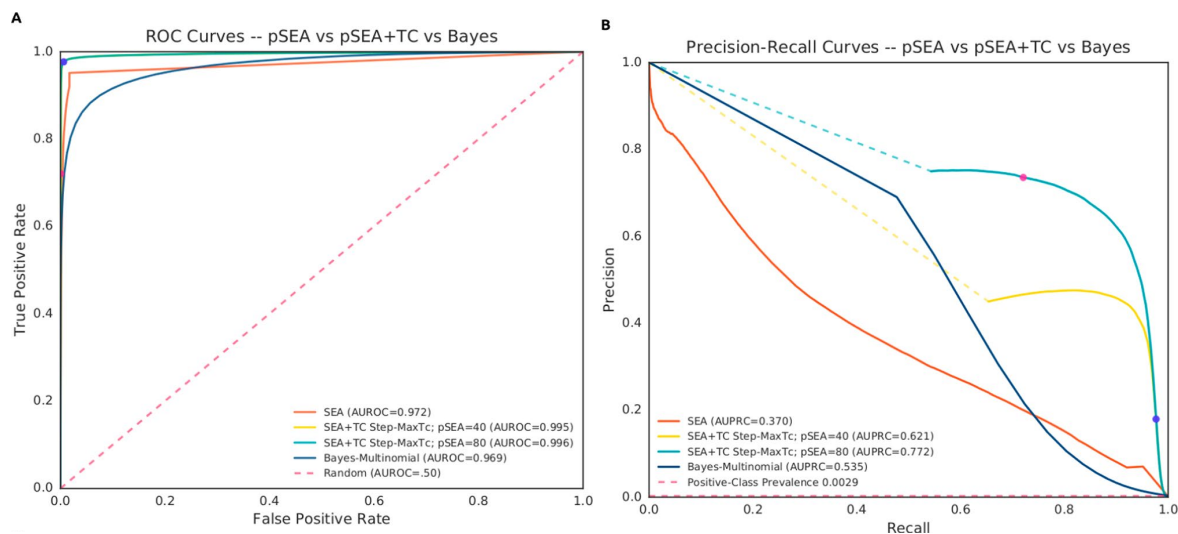


Figure 1.1 - Comparative performance of SEA, SEA+TC, and a multinomial naive-Bayesian classifier (NBC) on ChEMBL cross-validation sets. Receiver operating characteristic (ROC) curves from independent 5-fold cross-validation runs for each method. Methods are evaluated on independent cross-validation sets filtered for >5 ligands per ChEMBL protein target (equivalent analyses at >50 ligands per target reported in Supporting Information Figure S2). Overall performance is gauged by the area under the ROC curve (AUROC). Note, for SEA+TC cross-validation sets, ROC curves are the result of stepping a decision threshold across MaxTc values, while holding a separate pSEA decision threshold at 40 (yellow curve) or 80 (cyan curve) (see Methods). Complementary curves stepping across SEA p-values are available in Supporting Information Figures S1 and S2. Dotted lines span the distance between a fully stratified classifier (TPR = 0; FPR = 0) and the minimum point at which both SEA+TC decision thresholds begin to affect performance. Pink and blue circles indicate the recommended upper and lower bounds for MaxTc thresholding on their respective pSEA-threshold curves, respectively (upper = 0.80; lower = 0.40). (B) Corresponding precision-recall curves (PRCs) for cross-validation runs described in part A. Positive-class prevalence (dashed red line) indicates the chance of selecting a positive association from the data set at random (0.0014). Performance is measured by the area under the PRC (AUPRC).

(B)

A ZINC100 (Bucumolol)
In: anodyne bb for-sale in-man-only in-stock
Google Wikipedia PubMed

Heavy Atoms: 22

Added: 2005-11-15 Available: Premier Since: 2015-08-07 Mwt: 305.374 logP: 2.229 Tranche: DEAA

Activities based on ChEMBL 20
There is no known activity for this compound.

SEA Predictions based on ChEMBL 20 [Run SEA](#)

Gene	Description	Target Class	P-Value	Max Tc
ADRB1	Beta-1 adrenergic receptor	membrane receptor / GPCR-A	17	48
ADRB2	Beta-2 adrenergic receptor	membrane receptor / GPCR-A	33	44

B zinc15.docking.org/genes/ADRB2/substances/?ecfp4_fp-tanimoto-40=ZINC00000000100

ZINC2827950

0.50

ZINC3109233

0.50

ZINC3109234

0.50

ZINC3830339
Bunolol

0.50

Figure 1.2 - Predictions supported by evidence. (A) Here, Bucumolol (ZINC100) is shown with a SEA prediction for ADRB2 at a pSEA = 33 and MaxTc to the nearest annotated compound of 0.44. The user may click on the “44” to go to the URL shown, which lists bucumolol’s closest-match known ADRB2 ligands in decreasing order of similarity (the first four are shown). The user may also click on “Run SEA” to rerun a SEA calculation on the molecule, providing comprehensive statistics.

A

« 2 » 3311 / genes Filters SLC6

Name	Description	Organism	Sub Class (Major Class)	Purchasable	Predicted
ADH7	Alcohol dehydrogenase class 4 mu/sigma chain	Eukaryotes	enzyme-oth (enzyme)		3379
ADHFE1	Hydroxyacid-oxoacid transhydrogenase, mitochondrial	Eukaryotes	enzyme-oth (enzyme)		0

SLC6A11 Sodium- and chloride-dependent GABA transporter 3

SLC6A8 Sodium- and chloride-dependent creatine transporter 1

SLC6A9 Sodium- and chloride-dependent glycine transporter 1

SLC6A12 Sodium- and chloride-dependent betaine transporter

SLC6A5 Sodium- and chloride-dependent glycine transporter 2

SLC6A6 Sodium- and chloride-dependent taurine transporter

SLC6A1 Sodium- and chloride-dependent GABA transporter 1

SLC6A4 Sodium-dependent serotonin transporter

SLC6A15 Sodium-dependent neutral amino acid transporter B(0)AT2

SLC6A13 Sodium- and chloride-dependent GABA transporter 2

B

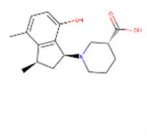
« 1 » 13 / genes Filters SLC6

Name	Description	Organism	Sub Class (Major Class)	Orthologs	Observations	Substances	Purchasable	Predicted
SLC6A1	Sodium- and chloride-dependent GABA transporter 1	Eukaryotes	electrochemical-transporter (Transporter)	3	588	336	17	1932338
SLC6A11	Sodium- and chloride-dependent GABA	Eukaryotes	electrochemical-transporter (Transporter)	2	34	23	4	71157

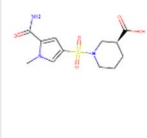
C

« 1 » 483487 / genes / SLC6A1 / predictions / subsets / strong Filters Lookup

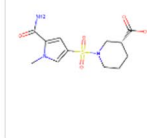
SLC6A1: ZINC226398813



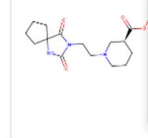
SLC6A1: ZINC37418492



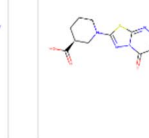
SLC6A1: ZINC37418494



SLC6A1: ZINC362355581



SLC6A1: ZINC49453787



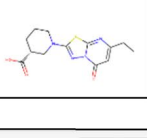
Applied Subsets

Strong : 386

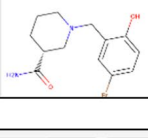
Other

- Dissimilar
- Identity
- Novel
- Purchasable
- Similar

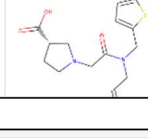
SLC6A1: ZINC49453788



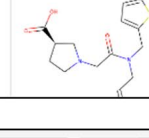
SLC6A1: ZINC154989283



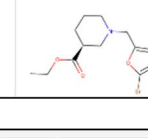
SLC6A1: ZINC62133856



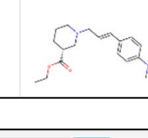
SLC6A1: ZINC62133857



SLC6A1: ZINC5326323



SLC6A1: ZINC4007745



D

« 1 » 234868 / genes / SLC6A1 / predictions / subsets / strong Filters Lookup

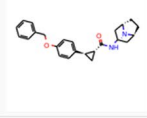
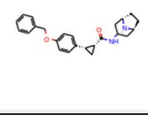
Gene	Description	Target Class	P-Value	Max Tc	Substance	Structure
SLC6A1	Sodium- and chloride-dependent GABA transporter 1	Transporter / electrochemical-transporter	169	45	ZINC000456322924	
SLC6A1	Sodium- and chloride-dependent GABA transporter 1	Transporter / electrochemical-transporter	169	45	ZINC000456322926	

Figure 1.3 - Tools to display predictions for a gene and filter and sort them by MaxTc and pSEA.

(A) Gene page showing predictions, with search bar to locate genes by name, top right. <https://zinc15.docking.org/genes>. (B) Gene listings for genes matching “SLC6” <https://zinc15.docking.org/genes/search?q=SLC6>. (C) Strongly predicted ligands for SLC6A11, showing the popup for subset selections <https://zinc15.docking.org/genes/SLC6A11/predictions/subsets/strong>. (D) Individual predictions, showing MaxTc and pSEA for each prediction, sorted by pSEA, with a MaxTc (novelty/similarity) limit specified <https://zinc15.docking.org/genes/SLC6A1/predictions/subsets/strong/table.html?sort=-pvalue&maxtc-between=40+45>.

4A

« 1 » 3311 /genes Filters GP 1

Name	Description	Organism	Sub Class (Major Class)	Predicted
KCNK3	Potassium channel subfamily K member 3	Eukaryotes	VGIC (ion channel)	7204004
DRD4	D(4) dopamine receptor	Eukaryotes	GPCR-A (membrane receptor)	6494008
KCNK9	Potassium channel subfamily K member 9	Eukaryotes	VGIC (ion channel)	5356070
CCR3	C-C chemokine receptor type	Eukaryotes	GPCR-A (membrane receptor)	5355276

GPI Glucose-6-phosphate isomerase

- GPT Alanine aminotransferase 1
- GPRC6A G-protein coupled receptor family C group 6 member A
- GPD1 Glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic
- GPR135 Probable G-protein coupled receptor 135
- GPR142 Probable G-protein coupled receptor 142
- GPR39 G-protein coupled receptor 39
- GPR81 G-protein coupled receptor 81
- GPR17 Uracil nucleotide/cysteinyl leukotriene receptor
- ART1 GPI-linked NAD(P)(+)-arginine ADP-ribosyltransferase 1

4B

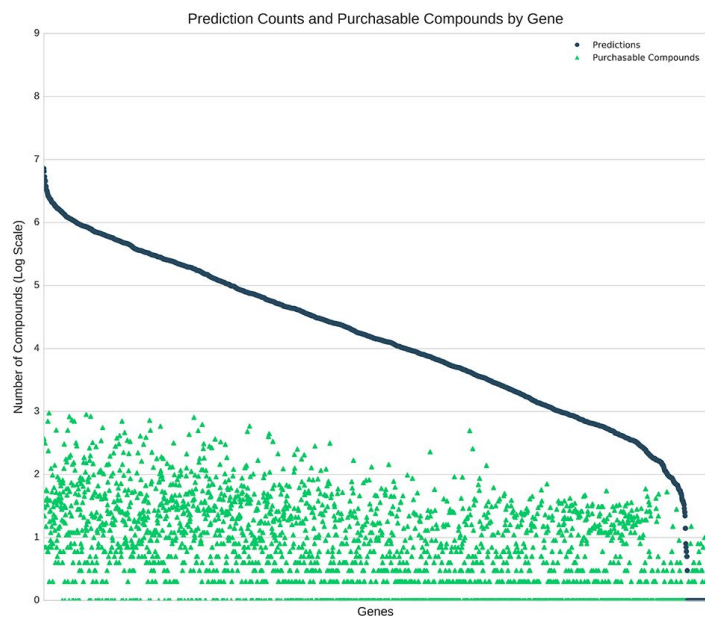
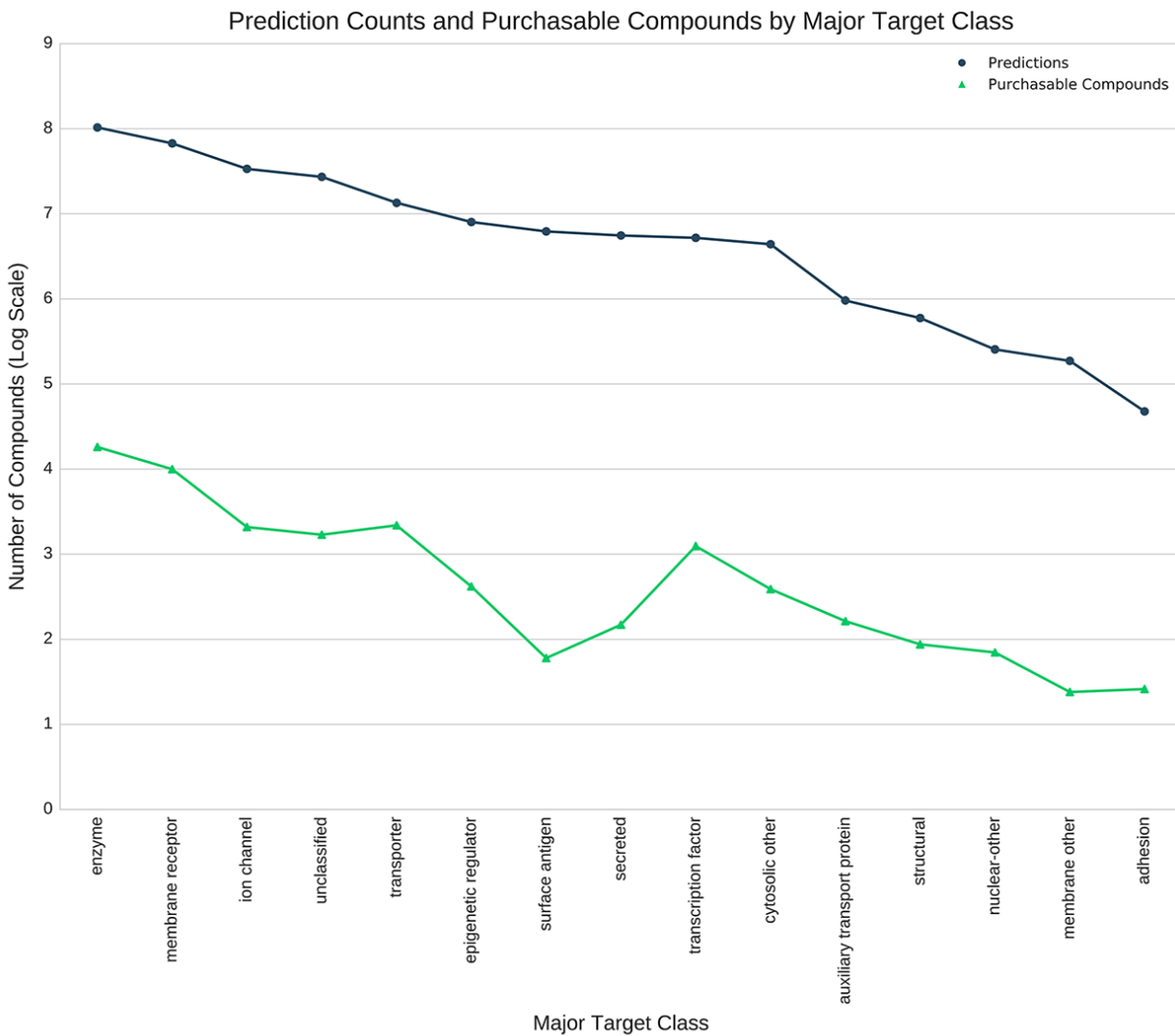
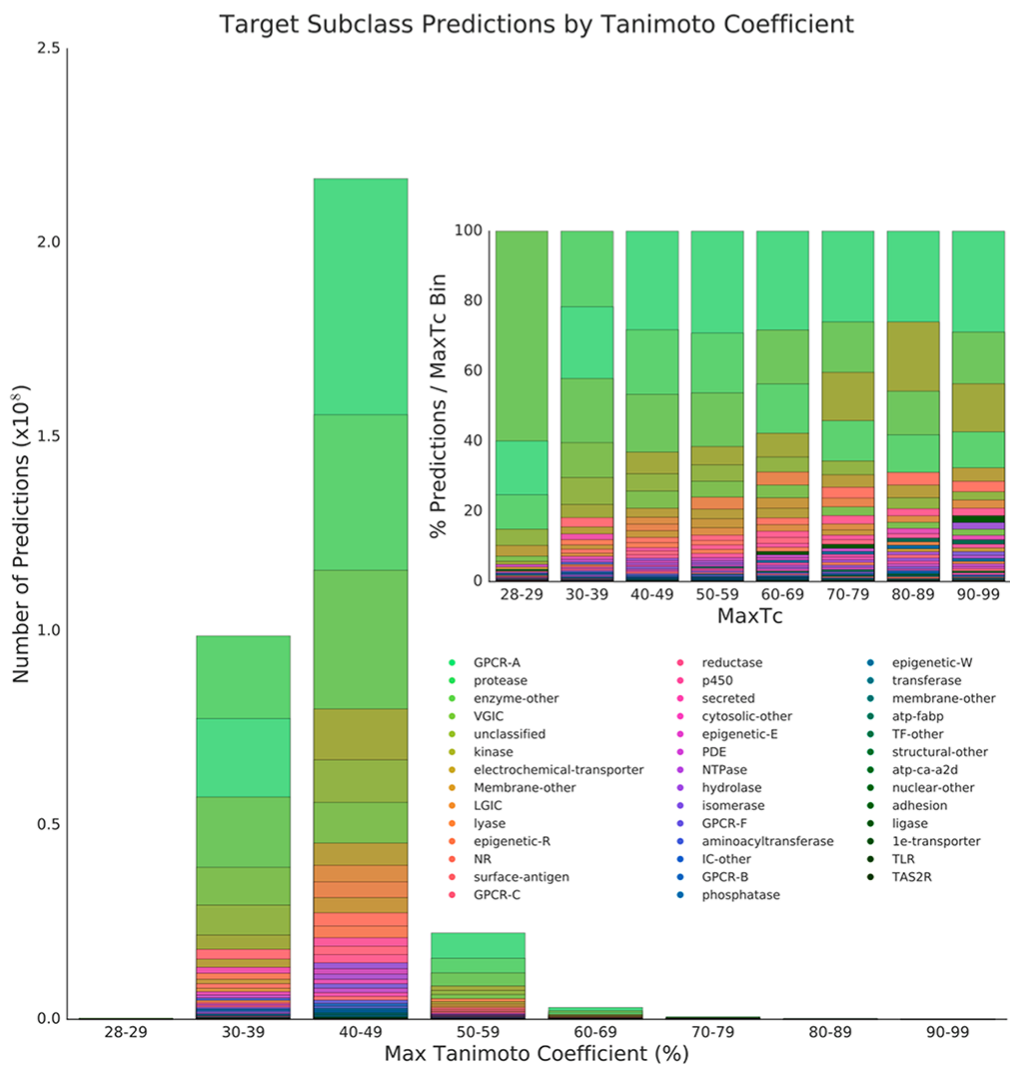


Figure 1.4 - Predictions available for 2629 genes. (A) The web interface allows genes and their predictions to be found by name or gene symbol: <https://zinc15.docking.org/genes>. Enter the gene name in the search field (1). Click on the predictions link (2) to display the predicted ligands. (B) Predictions and purchasable compounds for 2629 genes. The horizontal axis is genes, sorted by number of predictions. The vertical axis is number of compounds, log scale, labeled by exponent. Dark gray circles indicate the number of predicted purchasable compounds for a gene. Green triangles represent the number of purchasable annotated compounds for the same gene.

5A) By major target class. Data from <https://zinc15.docking.org/majorclasses>



5B). By target subclass. <https://zinc15.docking.org/subclasses>.



5C) By Kingdom, called organism class in ChEMBL and ZINC. Data from

<https://zinc15.docking.org/organisms>.

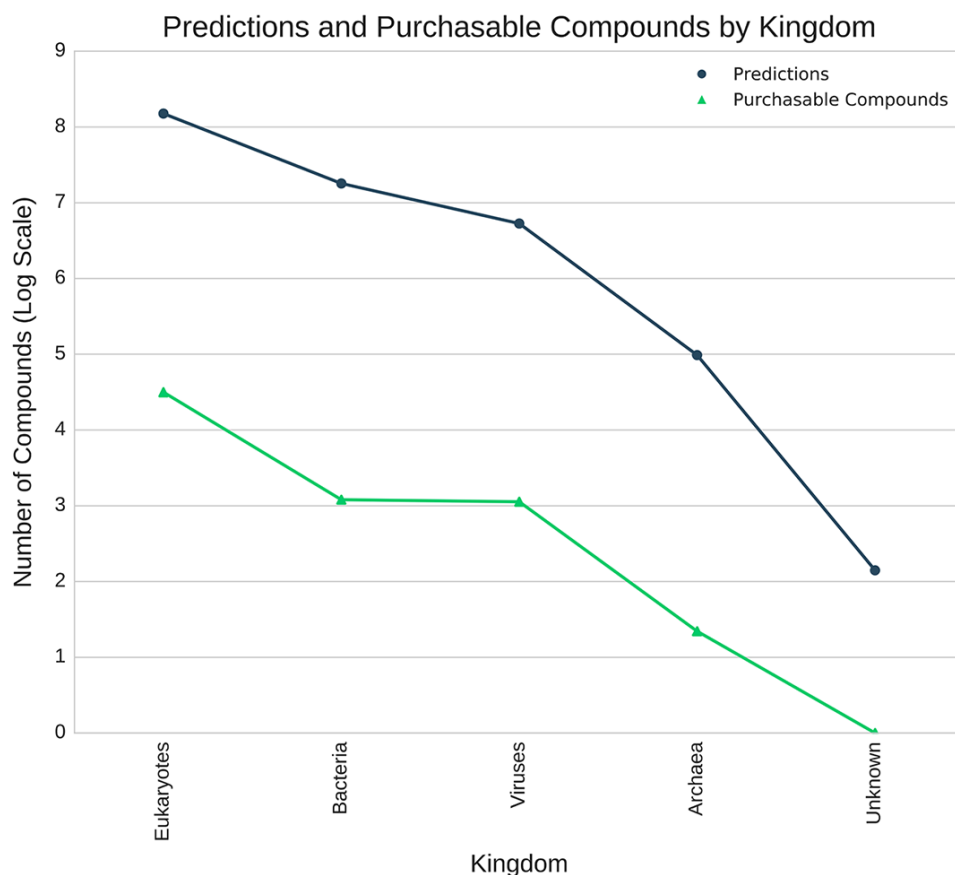
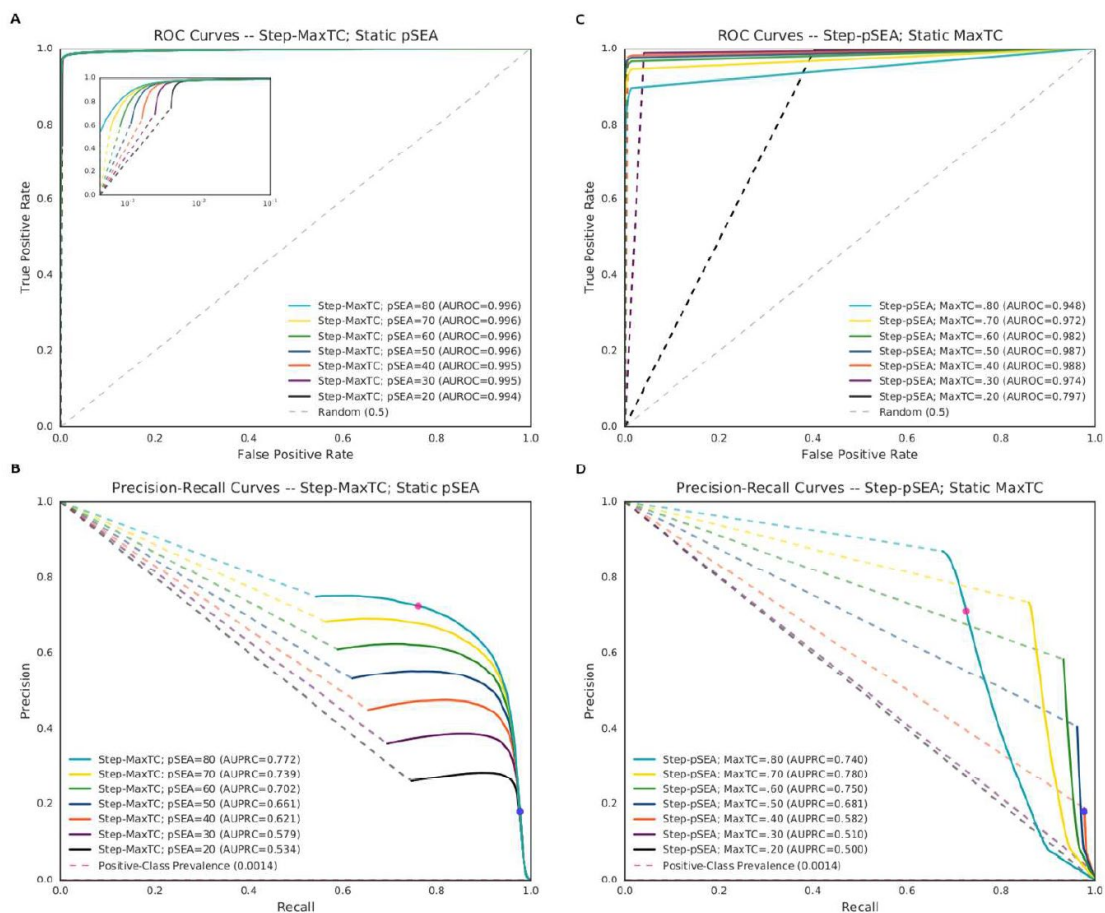
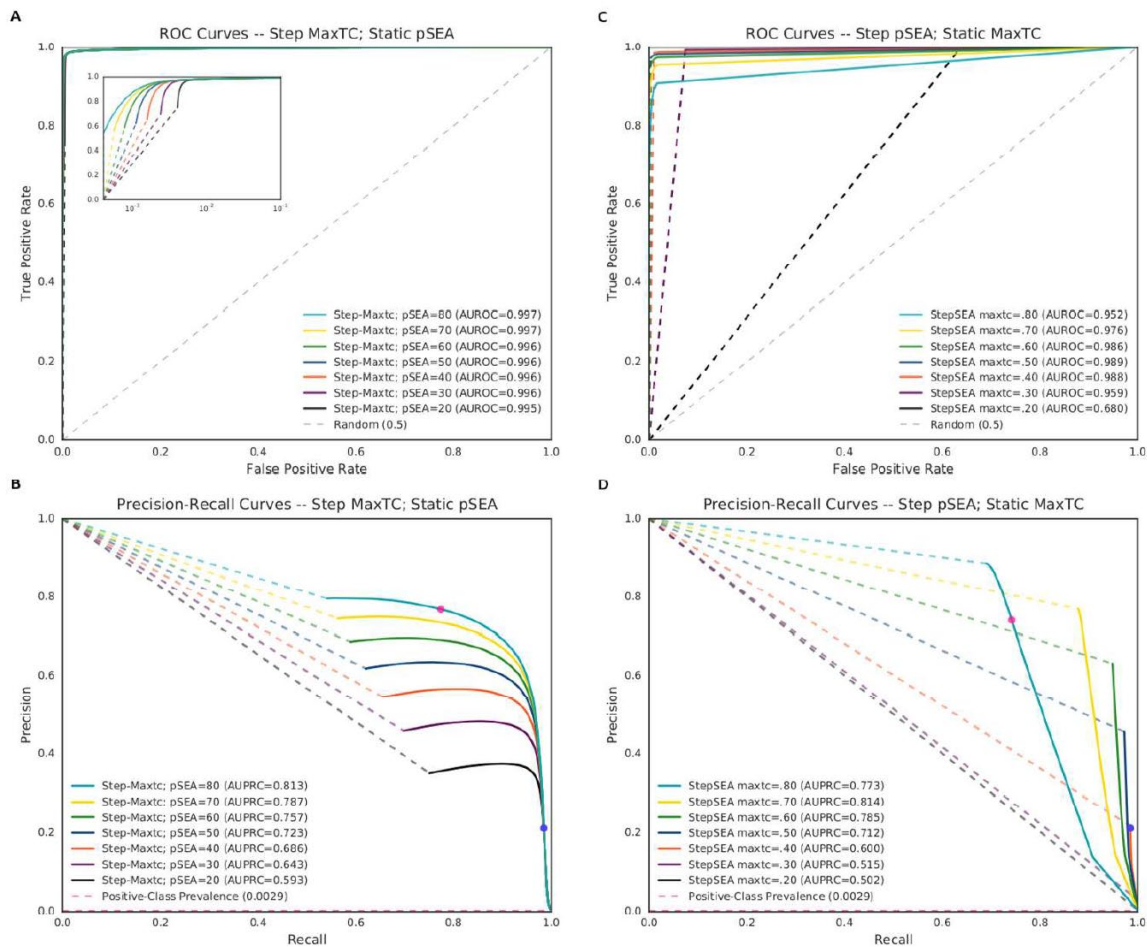


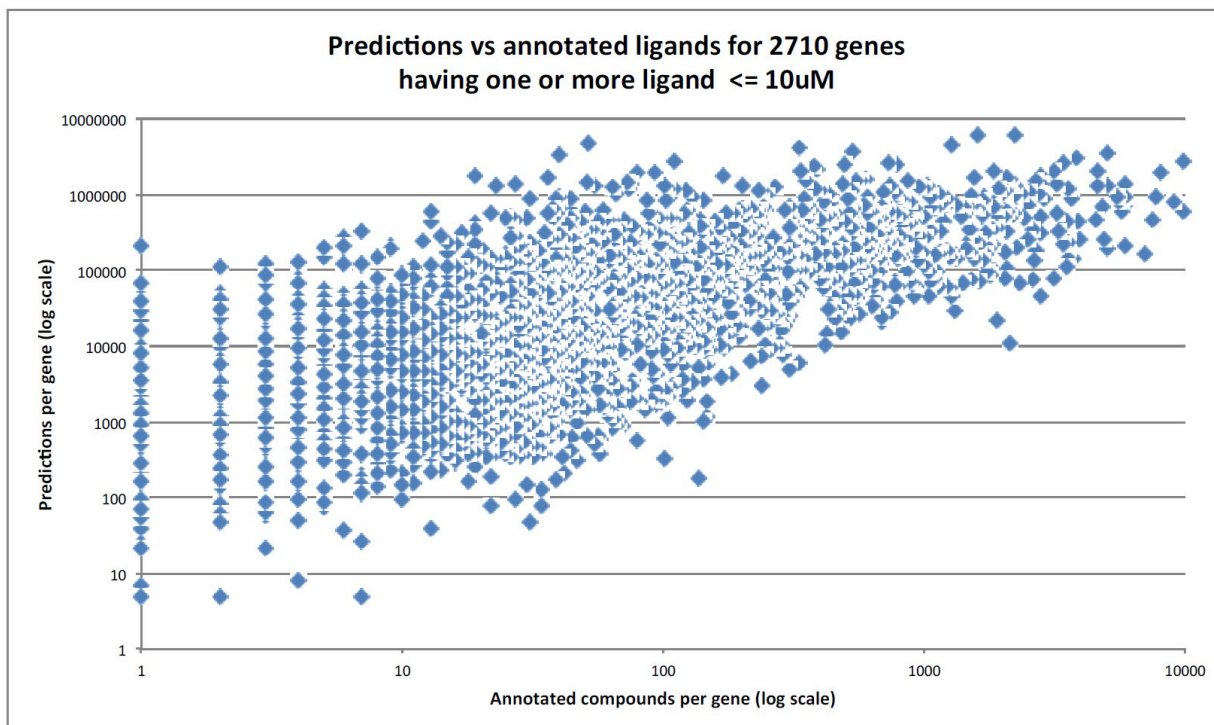
Figure 1.5 - Prediction counts and purchasable compounds. The gray line indicates the number of predictions, and the green line represents the number of annotated compounds. (A) By major target class. Data from <https://zinc15.docking.org/majorclasses>. (B) By target subclass. Most target predictions have a maximum tanimoto coefficient between 0.30 and 0.39 and 0.40–0.49. Percent of predictions for each target subclass relative to MaxTc are plotted in the inset to show the full spread of prediction across bins. (C) By Kingdom, called organism class in ChEMBL and ZINC. Data from <https://zinc15.docking.org/organisms>.



Supplementary Figure A.1.1 - Performance metrics for SEA+TC on ChEMBL cross-validation sets filtered for >5 ligand annotations per target. All curves are derived from independent 5-fold cross-validation runs. Overall performance is measured by either the AUROC (A, C) or the AUPRC (B, D). (A) ROC curves for ChEMBL cross-validation sets at a variety of threshold values. Each curve is the result of stepping the decision threshold across MaxTC values, while holding the SEA p-value decision threshold constant. Inset shows a zoomed-in version of ROC curves, with the FPR (x-axis) in log units to emphasize low-FPR behavior. (B) Corresponding PRCs for cross-validation runs described in (A). Pink and blue circles indicate the recommended upper and lower bounds for MaxTC thresholding, respectively (MaxTC = 0.80; 0.40). (C) Complementary ROC curves to section (A); each curve is the result of stepping across all SEA p-values, while holding the MaxTC decision threshold constant. (D) Corresponding PRCs for cross-validation runs described in (C). Pink and blue circles indicate the recommended upper and lower bounds for the SEA p-value decision threshold, respectively (pSEA = 80; 40).



Supplementary Figure A.1.2 - Performance metrics for SEA+TC on ChEMBL cross-validation sets filtered for >50 ligand associations per target. All analyses replicate those undertaken in Supplementary Figure A.1.1.



Supplementary Figure A.1.3 – Prediction bias. Plot shows predictions per gene vs annotated ligands per gene, highlighting a general trend toward more predicted ligands when more known ligands are available.

1.9 Tables

Table 1.1 - Drugs with No Binding Data in ChEMBL, Predicted by SEA or MaxTC, Corroborated by the Literature

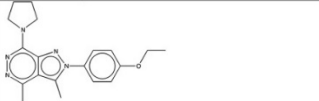
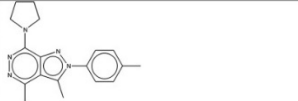
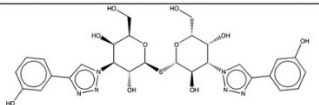
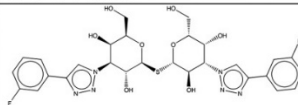
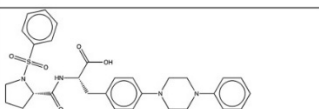
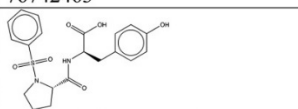
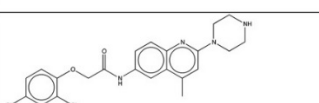
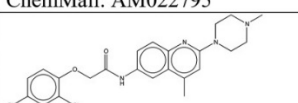
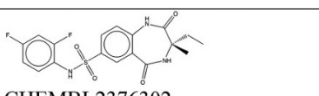
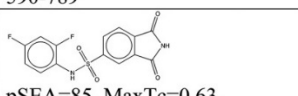
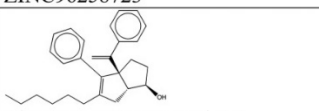

drug ^(ref)	ZINC ID	target	pSEA	MaxTc
Acemetacin ⁶³	601272	PTGS2	40	0.76
Afeletecan ⁶⁴	150339966	TOP1	69	0.41
Alclometasone ⁶⁵	4172330	NR3C1	15	0.58
Alminoprofen ⁶⁶	22	PTGS2		0.47
Amisulpride ⁶⁷	1846088	DRD3	22	0.66
Ancarolol ⁶⁸	39	ADRB2	42	0.44
		ADRB1	59	0.43
		ADRB3	29	0.44
Aranidipine ⁵³	600803	CACNA1C	121	0.75
		CACNA1D	132	0.51
Azasetron ⁶⁹	4132	HTR3A	25	0.61
Azelnidipine ⁷⁰	38141706	CACNA1C	91	0.56
		CACNA1D	124	0.57
Azetirelin ⁷¹	3804057	TRHR	95	0.59
		TRHR2		0.61
Besifloxacin ⁷²	3787097	PARC		0.46
Bevantolol ⁷³	1542891	ADRB1	89	0.51
		ADRB2	73	0.58
		ADRB3	73	0.53
Bilastine ⁷⁴	3822702	HRH1	48	0.51
Binospirone ⁷⁵	1999423	HTR1A		0.48
Bufetolol ⁵²	101	ADRB1	51	0.44
		ADRB2	47	0.46
Bunazosin ⁷⁶	601249	ADRA1B	52	0.61
Bupranolol ⁷⁷	106	ADRB2	45	0.44
		ADRB1	19	0.45
Butofilolol ⁷⁸	112	ADRB1	50	0.40
		ADRB2	34	0.46
Calcifediol ⁷⁹	12484926	VDRA		0.79
		GC		0.79
Camazepam ⁸⁰	2008504	GABARA5	25	0.53
		GABARA2	15	0.53

drug^(ref)	ZINC ID	target	pSEA	MaxTc
Cellcept ⁸¹	21297660	IMPDH1		0.70
		IMPDH2		0.70
Ciprokiren ⁸²	8214528	REN	178	0.68
Dasotraline ⁸³	2510873	SLC6A3	25	0.63
		SLC6A2	29	0.63
Demecarium ⁸⁴	3875376	ACHE		0.71
Dienesterol ⁸⁵	4742540	ESR1	26	0.46
		ESR2	15	0.46
Edaglitazone ⁸⁶	1483899	PPARG	83	0.66
		PPARA	83	0.65
Efonidipine ⁸⁷	38139973	CACNA1C	81	0.51
		CACNA1D	118	0.51
Eptazocine ⁸⁸	1846076	OPRD1	30	0.42
		OPRK1	30	0.46
		OPRM1	32	0.46
Etanterol ⁸⁹	263	ADRB1	23	0.47
		ADRB2	47	0.40
Ethylmorphine ⁹⁰	3629718	OPRD1	28	0.62
		OPRK1	24	0.62
		OPRM1	32	0.75
		OPRL1		0.57
Etomoxir ⁹¹	1851171	CPT1		0.47
Fiduxosin ⁹²	29747110	ADRA1A	30	0.53
		ADRA1B	45	0.53
		ADRA1D	38	0.46
Floxacillin ⁹³	4102187	BLAACC-4		0.80
Flurazepam ⁹⁴	537752	GABARA5	28	0.50
		GABARA1	17	0.49
Granisetron ⁹⁵	347	HTR3A	25	0.75
Halobetasol ⁹⁶	4214603	NR3C2	20	0.60
Hexoprenaline ⁹⁷	3872806	ADRB2	77	0.52
Ketobemidone ⁹⁸	1600	OPRD1	49	0.46
		OPRK1	45	0.48
		OPRM1	44	0.55
Lercanidipine ⁹⁹	19685790	CACNA1B		0.49
		CACNA1C	107	0.70
		CACNA1D	146	0.63

drug^(ref)	ZINC ID	target	pSEA	MaxTc
Lexacalcitol ¹⁰⁰	4474609	VDR	144	0.62
Meptazinol ¹⁰¹	854	OPRD1	44	0.48
		OPRK1	39	0.60
		OPRM1	38	0.55
Metipranolol ¹⁰²	494	ADRB1	27	0.45
		ADRB2	31	0.52
Ormeloxifene ¹⁰³	5104028	ESR1	86	0.51
		ESR2	58	0.44
Paroxypropione ¹⁰⁴	1890	ESR1	38	0.58
		ESR2	30	0.58
Pipenzolate ¹⁰⁵	601314	CHRM1		0.47
		CHRM2	30	0.43
		CHRM3	57	0.53
		CHRM4	35	0.53
		CHRM5	40	0.53
Pozanicline ¹⁰⁶	6562	CHRNA2	33	0.57
		CHRNA4		0.57
		CHRNA10	53	0.55
Propiverine ¹⁰⁷	1530934	CHRM2	24	0.42
		CHRM3	50	0.57
Revatropate ¹⁰⁸	4214265	CHRM1	55	0.53
		CHRM2	33	0.53
		CHRM3	59	0.57
		GPM3		0.57
Temazepam ¹⁰⁹	740	GABA5	28	0.59
Udenafil ¹¹⁰	13916432	PDE5A	74	0.61
Unoprostone ¹¹¹	8214703	PTGER1	45	0.57
		PTGER2	30	0.40
		PTGER3		0.57
		PTGDR	52	0.40
		PTGFR	85	0.51
Valategrast ¹¹²	72190226	ITGA4	60	0.32
Verubulin ¹¹³	35978229	TUBB3	62	0.51

Table 1.2 - Selected Plausible Predictions of Purchasable Compounds for Genes with No

Purchasable Ligands in ChEMBL

Gene symbol or UniProt code – Name – # annotated ligands	Annotated compound – ChEMBL code, pKi, ZINC ID (citation)	Predicted compound – pSEA, MaxTc, ZINC ID, Vendor: Vendor code
CACNA2D2 - Voltage-dependent calcium channel subunit alpha-2/delta-2, 26 ligands	 <p>ChEMBL1801206, pKi=7.7, ZINC72107844¹¹⁴</p>	 <p>pSEA=132, MaxTc=0.72, ZINC36646273, Specs: AO-476/43421055</p>
LGALS3 – Galectin-3, 38 ligands	 <p>ChEMBL2313626, pKi=7.66, ZINC95598439¹¹⁵</p>	 <p>pSEA=95, MaxTc=0.81, ZINC208938373, eMolecules: 76742463</p>
ITGA4 – Integrin alpha-4, 230 ligands	 <p>ChEMBL254140, pKi 7.12, ZINC28978676¹¹⁶</p>	 <p>pSEA=105, MaxTc=0.70, ZINC255966043, 1717-ChemMall: AM022795</p>
MCHR2 – Melanin-concentrating hormone receptor 2, 48 ligands	 <p>ChEMBL196667, pKi 8.43, ZINC13671957¹¹⁷</p>	 <p>pSEA=83, MaxTc 0.83, ZINC20114362, MolPort-007-590-789</p>
MOGAT2 – 2-acylglycerol O-acyltransferase 2, 38 ligands	 <p>ChEMBL2376302, ChEMBL2366303, pKi 8.00, ZINC96258723¹¹⁸</p>	 <p>pSEA=85, MaxTc=0.63, ZINC358638938, Enamine REAL: Z1143276235</p>
NR5A2 – Nuclear receptor subfamily 5 group A member 2, 50 ligands. Aka LRH-1.	 <p>ChEMBL1765959, pKi 6.6, ZINC71318097¹¹⁹</p>	 <p>pSEA=104, MaxTc 0.76, ZINC252079412, Ambinter: Amb22802160</p>

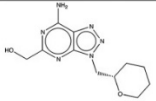
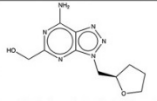
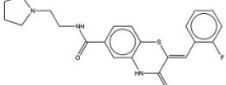
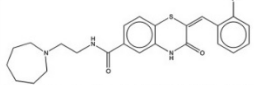
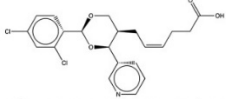
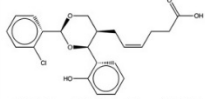
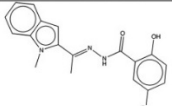
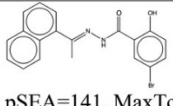
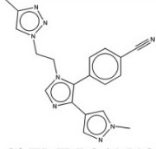
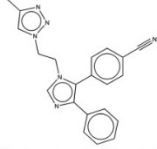
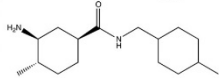
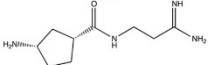
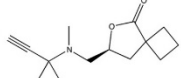
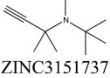
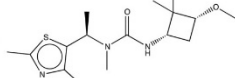
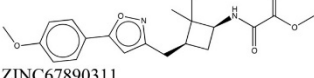
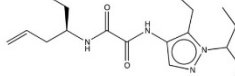
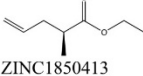
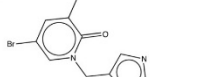
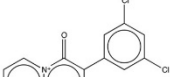
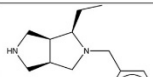

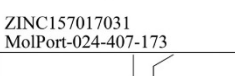
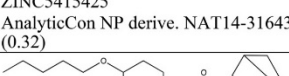
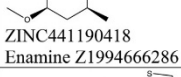
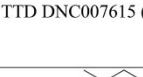
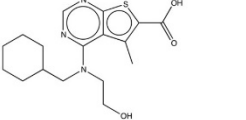
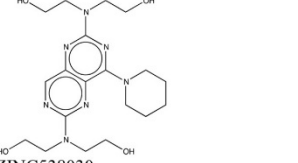
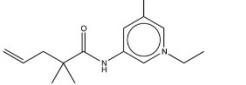

Gene symbol or UniProt code – Name – # annotated ligands	Annotated compound – ChEMBL code, pKi, ZINC ID (citation)	Predicted compound – pSEA, MaxTc, ZINC ID, Vendor: Vendor code
PDE8B – High affinity cAMP-specific and IBMX-insensitive 3',5'-cyclic phosphodiesterase 8B 184 ligands	 ChEMBL2070904, pKi 6.11, ZINC84688179 ¹²⁰	 pSEA=136, MaxTc = 0.93, ZINC40450041, eMolecules: 30177261
PFG6PD – Glucose-6-phosphate dehydrogenase, 33 ligands	 ChEMBL2170912, pKi 5.02, ZINC37492541 ¹²¹	 MaxTc 0.98, ZINC20734607, eMolecules: 27208436
PTGIS – prostacyclin synthase, 64 ligands	 ChEMBL355596, pKi 5.6, ZINC13740452 ¹²²	 pSEA=0, MaxTc=0.58, ZINC3782896, Tocris: 0837
PYK – pyruvate kinase, 58 ligands	 ChEMBL2206713, pKi 6.42, ZINC254133200 ¹²³	 pSEA=141, MaxTc=0.64, ZINC6312342, AsisChem: W72945
BAZ2A, Bromodomain adjacent to zinc finger domain protein 2A, 7 ligands	 ChEMBL3415182, pKi=6.22, ZINC263621294 ¹²⁴	 MaxTc=0.82, ZINC65410133, ChemBridge: 56126287

Table 1.3 - Compounds with No Predictions: “Chemical Dark Matter”

Chemical Dark Matter ZINC ID Vendor: Catalog item	Nearest bioactive ZINC ID Annotation (MaxTc to dark matter)
 ZINC221102357 AKOS006188593 MolPort-014-310-839	 ZINC4214812 Streptom: 3114 (0.29)
 ZINC491727059 Enamine Z2063391489	 ZINC31517377 Butynamine, an antihypertensive (0.28)
 ZINC440247506 Enamine Z1375141689	 ZINC67890311 Analyticon NP derivative NAT26-505611 (0.32)
 ZINC440251478 Enamine Z1992115809	 ZINC1850413 HMDB29761 Fema3489 (flavor) (0.29)
 ZINC156966658 MolPort-031-342-219	 ZINC207174299 CHEBI:131599 Mesionic insecticide (0.31)
 ZINC157017031 MolPort-024-407-173	 ZINC5415425 AnalyticCon NP derive. NAT14-316437 (0.32)
 ZINC441190418 Enamine Z1994666286	 ZINC28712056 TTD DNC007615 (0.31)
 ZINC448649310 Enamine Z2254466482	 ZINC456 DrugBank DB00961 (0.32)

Chemical Dark Matter ZINC ID Vendor: Catalog_item	Nearest bioactive ZINC ID Annotation (MaxTc to dark matter)
 <p>ZINC572486678 Enamine Z2465963117</p>	 <p>ZINC538030 MOPIDAMOL, a PDE inhibitor (0.30)</p>
 <p>ZINC450432369 Enamine Z2272296490</p>	 <p>ZINC2236 Valdetamide – a sedative (0.32)</p>

Chapter 2: Automating Diagnosis of Melanocytic Atypia: A Precursor to Melanoma in Situ

Gaskins G, Mew N, Keiser MJ⁺, Keiser E⁺

University of California San Francisco, Institute of Neurodegenerative Diseases, 675 Nelson Rising Lane,
San Francisco, California 94143, United States

⁺ Corresponding authors

2.1 Abstract

Melanocytic atypia is histopathologically challenging. Pathologist interobserver agreement for melanocytic atypia in standard (H&E) histology images is low, ranging from 33-68%, with melanoma in situ (MIS) in particular contributing to diagnostic discordance. A lack of agreement among experts presents a challenge to any supervised learning task, where the utility of a learned function depends on the accuracy and reliability of labels used. To circumvent the issue of discordance in human labeling, we pair H&E histology images with contiguously cut tissue sections, immunohistochemically (IHC) stained for melanocytes via one of three antibodies: Melan-A, SOX10, or MITF. We develop a deep-learning pipeline for automating diagnosis of melanocytic atypia using a custom dataset of 172 (81 H&E, 81 IHC) paired, whole slide images (WSIs). Networks are trained to identify the presence of melanocytic atypia in H&E sections using information from paired samples possessing either one (Melan-A; SOX10) or multiple (Melan-A + SOX10: MESX) IHC stains. Networks trained on sample pairs possessing a single IHC stain achieve strong performance on a holdout patient dataset, as assessed by the area under the receiver-operating characteristic (AUROC) and area under the precision-recall curve (AUPRC), respectively (Melan-A: 0.940, 0.836; SOX10: 0.901, 0.831). To visualize model predictions, we generate full-scale (20X magnification) high-resolution (pixel-wise) prediction heatmaps on holdout tissue sections (H&E) for pathological interpretation. Additionally, saliency mapping shows what features activate the network most, and that saliency profiles vary based on IHC stain used to extract labels. This finding indicates different antibodies may provide alternate but complimentary information for solving this classification problem. We hope this pipeline will provide assistance to the clinical pathologist to reach better consensus regarding new MIS diagnoses in cutaneous biopsies.

2.2 Introduction

2.2.1 Deep Learning in Medical Image Analysis

Recently, deep learning has made strides in the field of medical image analysis¹²⁵⁻¹²⁷. Convolutional neural networks¹²⁸ have proven to be effective at identifying and diagnosing a wide range of pathologies¹²⁹, including skin disease¹³⁰⁻¹³³, and diagnostic performance on trained networks approaches, or in some cases surpasses, that of the average pathologist¹³⁴⁻¹³⁵. Most of these achievements rely on a paradigm of supervised learning, in which networks are trained from a corpus of well labeled, human curated images. Though new techniques are emerging to alter the resolution of labeling required¹³⁰, for the majority of cases, a pathologist must manually provide ground-truth labeling. If there is little consensus regarding what the ground-truth label should be, the classification task becomes more difficult and the model less generalizable.

2.2.2 Melanocytic Atypia is Histopathologically Challenging

Such is the case with melanocytic atypia, an early stage precursor to melanoma that is histopathologically challenging to identify. Pathologist interobserver agreement for melanocytic atypia in standard (H&E) histology images is low, ranging from 33-68%, with melanoma in situ (MIS) in particular contributing to diagnostic discordance¹³⁶⁻¹³⁷. A lack of agreement among experts presents a challenge to any supervised learning task, where the utility of a learned function is dependent upon the accuracy and reliability of the labels used to train it. To circumvent the issue of diagnostic discordance negatively contributing to our ground-truth labels, we apply a novel, pathologist-agnostic method to identify melanocytic atypia in patient tissue sections.

2.2.3 Circumventing the Necessity for Pathologist Manual Labeling

Our method utilizes tissue sections immunohistochemically (IHC) stained for melanocytes to create and label a corpus of corresponding H&E tissue sections for CNN training (see Methods). While IHC staining of patient biopsies is less common, additionally invasive, and more expensive than standard H&E staining, it provides a biomarker detailing where melanocytes reside. This biomarker is useful for identifying melanocytic atypia, as melanocytic atypia is inherently restricted to where melanocytes exist. However, due to the low prevalence of IHC stains within the clinic as well as the limitations mentioned above, it is impractical to train a CNN on IHC-tissue samples directly. To account for this, we assemble a custom dataset comprised of adjacently sliced and paired tissue sections. Each set of paired tissues consists of one slice that is IHC-stained and a complimentary slice that is H&E-stained. We refer to both slices, collectively, as a “sample pair”.

The IHC-stain information from each sample pair is extracted and serves to label that pair’s H&E image. This allows the method to incorporate information regarding melanocyte locations while still training a CNN on (labeled) H&E data. Ideally, the CNN learns features indicative of melanocytic atypia that are generalizable to new H&E images. Figure 2.1 details the methodology pipeline. This method relies on a critical assumption: that contiguously sliced tissue sections are similar enough in morphology and location that IHC staining from one slice can serve as a proxy label for its adjacent (H&E) pair. To maximize this assumption and to ensure optimal morphological congruence among paired H&E and IHC images, we align sample pairs at their native 40X resolution (see Methods). We then train a convolutional neural network (CNN) to identify melanocytic atypia in H&E images, at the tile level.

We assess performance of our (H&E) trained networks by their ability to classify tiled images from holdout H&E patient tissue sections as either melanocyte-containing, or non-melanocyte-containing, using the holdout’s unseen IHC-stained tissue as ground truth. To provide pathological interpretation, we generate full-scale (20X magnification) high-resolution (pixel-wise) prediction heatmaps on holdout

tissue sections (H&E). We also perform saliency mapping to highlight which image (tissue) regions activate the network most.

Teaching a supervised model to learn where atypical melanocytes are present in an H&E slice, solely from information from a corresponding IHC section, helps to reduce the necessity, cost, and time of manual dermatopathologist labeling, and circumvents the issue of noisy labeling¹³⁶⁻¹³⁷. We hope this method will assist the clinical pathologist in reaching better consensus regarding new MIS diagnoses for cutaneous biopsies.

2.3 Methods

2.3.1 Ethics Approval

All materials used are obtained either through pre-approved teaching slides from UCSF's or Stanford's dermatology programs, or from patients that gave informed consent for the distribution and use of their samples. Approval is overseen by the Institutional Review Board (IRB) at the University of California, San Francisco and Stanford University. Access to data follows current laws, regulations, and IRB guidelines. Patient samples used (WSIs) are de-identified, and do not contain any personal health information.

2.3.2 Sample Cohort

Patient histology sections are obtained from either UCSF's Dermatopathology and Oral Pathology Service or the dermatopathology service at Stanford's Anatomic Pathology and Clinical Laboratories. All samples obtained across both institutions are originally gathered between 2011-2015, with patients ranging in age between 43-73. Each patient sample contains a pair of WSIs: the first WSI consists of H&E-stained tissue sections and the second contains corresponding IHC-stained sections.

Samples are extracted from a variety of biopsies, including excisional/incisional, shave, and punch biopsies, but each pairing of H&E and IHC WSIs is extracted from the same patient biopsy.

2.3.3 Sample Preparation

Tissue sections from patient biopsies are sequentially sliced and prepared for one of two conditions: H&E staining or IHC staining. For H&E staining, patient samples are formalin fixed, paraffin embedded, and stained with hematoxylin and eosin (H&E). For IHC staining, tissue samples are similarly prepared, but treated with one of four antibodies for staining melanocytes: Melan-A, SOX10, MITF, or MelPro (Melan-A+Ki-67). Sample preparation alternates between H&E and IHC staining after each slice to ensure regional similarity between contiguous, paired sections. All H&E and IHC sections from a biopsy are placed onto two separate glass slides and scanned to generate a pair of WSIs. Whole slide images are digitized at 40X magnification, corresponding to a resolution of 0.25 microns per pixel (MPP), using an Aperio scanscope (AxioVision; Leica Biosystems) and stored as compressed pyramidal TIFF files (.SVS images).

2.3.4 Dataset Split

Table 2.1 (SOX10), Table 2.2 (Melan-A), and Table 2.3 (MITF) detail the composition of our histology dataset, which consists of 81 patient sample pairs (172 WSIs), including 690 individual tissue sections. Each sample pair is composed of a WSI containing H&E-stained histology sections, and an independent but complimentary WSI containing IHC stained histology sections (see above). Patient samples from both institutions (UCSF: 41/81 samples; Stanford: 40/81 samples) were grouped by their IHC stain to form three training sets: A “Melan-A” training set, composed of samples possessing either Melan-A or Melpro IHCs (combined: 32/81 samples), a “SOX10” training set (30/81 samples), and an “MITF” training set (19/81 samples), with the latter two consisting of samples containing their respective IHCs. We train three stain-specific convolutional neural networks (CNNs) using the subset of samples

from each training set (Melan-A: 15/32; SOX10: 12/30; MITF: 8/19) that pass all preprocessing stages (extraction, filtering, and alignment; see below). While the CNN trained on MITF samples does not perform well enough to use for prediction purposes, the MITF samples are used to help calibrate thresholds for stain extraction (see below; **Figure 2.2**). To incorporate multiple stains into a single model, we train an additional hybrid model using the combined set of Melan-A and SOX10 training samples (27/62).

2.3.5 Tissue Extraction and Filtering

To automate tissue extraction, we use a custom toolkit created in Python (**Supplementary Figure A.2.1**). The toolkit first identifies the HSV color-space of a WSI's background. Using this information, it then selects candidate tissue regions in the foreground that meet a size requirement. Thresholds for chroma, size, and degree of region-enclosing are manually adjustable, allowing for optimal tissue extraction. Once tissue regions are refined for both the H&E WSI and its corresponding IHC WSI, tissue sections from each pair are matched. Matching occurs automatically based on tissue locations in the WSI and can be manually adjusted by the user if necessary. Cases where either the matched H&E or IHC section are digitally corroded, have excessive tissue damage, are missing, are too small, or are indistinguishable from background, are filtered out. Post matching and filtering, successfully extracted tissue sections are individually cropped and background masked. Our toolkit employs the open-source libraries Pyvips¹³⁸, OpenCV¹³⁹, and IpyWidgets¹⁴⁰.

2.3.6 Stain Threshold Calibration

The design of our study incorporates multiple IHC stains that generate different biomarkers of melanocyte presence. The presence of these biomarkers dictates the labels a tile receives. Therefore, it is important to set decision thresholds that retain similar boundaries for positive and negative classes, independent of stain type or stain location (e.g., nucleus, cytoplasm). To ensure equity in thresholding

choices, we rely on the assumption that the average number of melanocytes across a sufficiently large number of tissue sections should fit a normal distribution. Under this assumption, we test and calibrate thresholds for each stain type, such that when the decision threshold is applied, the resulting distribution of positively labeled tiles for one stain, is no different from the distribution of positively selected tiles for another stain. To measure distribution similarity across stains, we perform, for each threshold, a Mann-Whitney U-Test for all stain combinations (**Figure 2.2**). We conduct Mann-Whitney U-Tests for all thresholds in the range of 0.4 to 6.0, with a step size of 0.1, and optimize for the thresholds that maximize similarity melanocyte distributions across all stain combinations. The optimal thresholds identified (Melan-A: 3.9%; SOX10: 0.8%, MITF:3.7%) are used for all analyses in this body of work.

2.3.7 Image Alignment and Tiling

A key feature of our analysis utilizes IHC stained tissue sections as proxy labels for matched H&E sections. This method rests on the assumption that contiguous slices are morphologically similar, and that a proxy label from one slice can relay information about the other (i.e., at the local level, the density of IHC staining for melanocytes is correlated with the probability of melanocytes being present within the same region of the preceding H&E slice). As feature extraction during model training is assessed at the pixel level, alignment between H&E and IHC tissue sections ensures optimal morphological congruence.

Matched H&E and IHC tissue sections are aligned at the slice level, at native resolution (40X magnification; 0.25 MPP), using the open-source alignment algorithm bUnwarpJ¹⁴¹. Briefly, the algorithm performs image registration using 2D elastic deformations (B-splines) to minimize the difference between two images, as calculated by an energy function. Matching tissue sections often vary in shape and size after extraction from WSIs. To align images properly, the smallest section from each matched pair is 0-padded to the size of the larger section before alignment. Additionally, matched sections too large to align (>5GB) are cropped to create 2 smaller pairs in place of one large pair.

Alignments are manually curated and given a score by the user, in addition to assessment by coefficients returned from the alignment algorithm. Alignment scores and manual scores exhibit a similar trend (**Supplementary Figure A.2.5**). Sample pairs that are unalignable or receive a low alignment score are discarded from the dataset.

Following alignment, matched H&E and IHC tissue sections are down sampled to 20X magnification and both images are identically partitioned into tiles of 256x256 pixels. Tiles with more than 80% of their pixels accounting for background are discarded. Additionally, several regions of tissue are marked in blue ink to delineate damaged tissue by an expert pathologist. Tiles containing blue ink (HSV color range [98, 56, 74], [116, 199, 252]) are removed from analysis if the ratio of foreground pixels to blue tinted pixels is less than 1.5X.

2.3.8 Denoising IHC Sections

Several tissue sections contain regions of damaged tissue and discoloration. This discoloration, at times, overlaps with the pigmentation of our IHC stain, thereby contributing false positives to our ground-truth IHC signal. We sought to denoise our dataset to remove these false positive regions. As tiles containing noisy, damaged tissue are easily identifiable to the human eye, we set up a web application through Amazon Web Services, similar to Tang et al., 2019, that allows users to label tiled images as either “noisy” or “not-noisey”. The web application is fed positively labeled IHC tiles from tissue sections harboring noisy regions, and the user is shown the tiled images along with tile location within the tissue. A graduate student familiar with the IHC stains in question labeled ~6000 tiles as either noisy (2631) or not-noisy (3639). Using these labeled images, we train a separate CNN to identify noisy tiles across all tiles in the dataset. Tiles highly predicted to contain noise ($>.80$) are removed from the dataset before training.

2.3.9 Performance Assessment

A substantial portion of the dataset was discarded during the preprocessing stages, forcing us into a lower data regime. As such, we could only afford to exclude a single patient from training for each of the stain-specific models. For each holdout patient, the combined set of tissue sections from the respective WSIs (Melan-A: 5/6 sections; SOX10: 3/3 sections) constitutes the holdout for each stain. Performance is assessed on the network's ability to predict H&E tile labels derived from the unseen IHC counterpart, for all tissue samples that comprise the holdout patient dataset. We measure performance by the AUROC and AUPRC for classification of tiles. To account for bias in selecting only one holdout patient, we conduct a leave-one-out analysis, that would ordinarily be a supplementary figure, but for sake of this thesis deadline, and as this analysis is not finalized, it is not included in this body of work. In order to reduce the substantial computational time a full round-robin leave-one-out analysis would take (>2 months), we reduce the ratio of negatives examples used for training from 10X to 3X for each iteration of the analysis.

2.3.10 Prediction Heatmaps

Prediction heatmaps (**Figure 2.3, Supplementary Figure A.2.2**) provide a visual representation of the CNN's confidence in melanocyte presence across the entirety of an H&E tissue section. Heatmaps are generated at the full 20X resolution using a custom multiprocessing script in combination with pytorch¹⁴². To generate predictions, every 6th pixel in the image serves as the center to a 256x256 pixel window. Windows are zero-padded to reach the appropriate size when near the edges of the image. Each window is passed to the trained CNN to obtain a prediction score. Prediction scores are converted and normalized to RGB values in the viridis color space, and filled into the original image using python. This process is done independently for each trained CNN (Melan-A; SOX10; MESX). Each heatmap generated involves millions of predictions, which we are able to process in the span of only few hours using a small number of GPUs.

2.3.11 Saliency Mapping

Saliency mapping methods attempt to provide intuition regarding what features a CNN finds important about an image. To examine which aspects of H&E tissues sections are differentially activating, we performed the guided gradients and integrated gradients techniques, as provided by <https://github.com/utkuozbulak/pytorch-cnn-visualizations>. Supplementary Figure A.2.4 shows an example where tissue sections of the slice are differentially activated (preferred) by the network.

2.4 Results

2.4.1 Stain Calibration

Labeling tiles requires setting a pixel threshold on the amount of IHC stain necessary to assign a (H&E) tile as positive. Because the design of our study incorporates multiple IHC stains, and therefore different biomarkers for melanocytes, it is important to set decision thresholds that retain similar boundaries for positive and negative classes, independent of stain type or stain location (see Methods). We find that the thresholds that maximize the similarity of distributions by U-Test comparison scores (max-score; most similar = 1.0) are 3.9% for Melan-A staining, 0.8% for SOX10, and 3.7% for MITF (**Figure 2.2B**). We use these thresholds for all analyses in the study.

Additionally, we cross predict on datasets using CNNs trained on data labeled from a stain different than the dataset under evaluation (and therefore, labeled under a different threshold). As all information passed to CNNs come from H&E images, cross-predictions highlight the biases inherent to the dataset a CNN was trained on, as well as whether the features learned from one form of labeling overlap or generalize to the ground-truth of another form of labeling. Cross-prediction performance scores for independently trained Melan-A and SOX10 CNNs perform well, albeit worse than on their own datasets (**Figure 2.2C**), achieving respectable classification performance values, while taking a significant

hit to precision-recall scores (Melan-A vs SOX10; 0.930, 0.617; SOX10 vs Melan-A, 0.912, 0.576 as determined by the AUROC and AUPRC, respectively).

2.4.2 Melan-A Model

We train a CNN on tiled images labeled from sample pairs containing Melan-A or MelPro IHC-stains. Tiled images consist of ~68,000 positive tiles and ~170,000 negative tiles derived from tissue sections. As negative examples from H&E samples are morphologically different, and therefore potentially trivial to learn to separate, we also inject an additional 170,000 negative tiles that are randomly sampled from tissue sections derived from TCGA's WSI image database. These negative TCGA injections serve to improve generalization, and to ensure our network does not simply evolve to be a "purple stain finder". To visualize TCGA tissue diversity and to interrogate model accuracy after training, predictions on a holdout subset of 50,000 randomly selected TCGA tiles are designated as either false-positive or true-negative, based on a prediction threshold of .50. Of the 50,000 tiles predicted, only 21 tiles (0.00042%) were False Positives (**Supplementary Figure A.2.3**), indicating the CNN is robust to morphologically similar images.

We examine performance on a holdout patient consisting of 5 separate tissue sample pairs. Performance is assessed by the AUROC and AUPRC, for the classification task of identifying the presence of melanocytes (via presence of stain). Ground-truth is determined by labels derived from the paired, IHC-stained slice. The Melan-A model achieves respectable performance metrics of 0.940 AUROC and 0.831 AUPRC on the holdout patient samples (**Figure 2.3A**). Whole tissue prediction heatmaps appear visually similar to the unseen IHC-stained images, despite never having direct access to this information (**Figure 2.3B**).

All tile predictions are assigned truth-table values based upon the comparison of the predicted label to the ground-truth label. Predicted labels are assigned based on a prediction threshold of .50 (>.50: positive prediction). Figure 2.4A shows randomly selected cases from each class. We observe that several

images falling into the False Negatives category have blurry H&E inputs (**Figure 2.4B**), potentially contributing to the error in prediction.

To visualize what the model found salient, we perform guided gradients on a strip of tissue comprised of seven overlapping tiles (**Supplementary Figure A.2.4**), stitch the tiles together, and color pixels according to the outcome by which they are most activated. Blue/purple regions indicate pixels have more activations when the prediction is positive, while red/green regions are activated more when the prediction is negative. Together, this image shows areas of the tissue with cellular morphologies are activated most when the image is predicted to be positive.

2.4.3 SOX10 Model

Similar to the Melan-A model above, we train a CNN tiled images labeled from sample pairs containing SOX10 IHC-stains. Tiled images consist of ~48,000 positive tiles and ~1430,000 negative tiles derived from tissue sections. An additional ~143,000 negative tiles from TCGA were injected, for reasons clarified in section 2.4.2. We assess performance of this model on a single patient holdout, consisting of three separate tissue sample pairs. Initial performance metrics were lower than expected, as this staining mechanism is often regarded in the field as more specific to melanocyte location. Upon further examination, we found several of the sample pairs used for training contained large regions of False Positives within our ground-truth labels, due to noise from tissue damage. Following our denoising procedure (see Methods), performance for the SOX10 network improved, with metric scores near that of the Melan-A model (AUROC: 0.907, AUPRC: 0.831, **Supplementary Figure A.2.2**). Whole tissue prediction heatmaps for SOX10 H&E holdout sections also elicited similar profiles to their unseen IHC-stained counterparts.

2.5 Discussion

The method we develop attempts to address a problem at the heart of all supervised learning techniques – that outputs are only as good as their inputs. Typically, this is in reference to the saying, “trash in, trash out”, when data quality is poor. Conversely, the idea applies to the other end of the spectrum as well. Training on the highest quality data available achieves the very best results possible. The real world issue arises when the very best result, say for the task of identifying melanocytes, remains unsatisfactory, due to humans (even professional ones) being fundamentally error-prone to the task at hand.

Rather than champion human subjectivity as an appropriate arbiter, we attempt to solve our task by exploiting biology as a surrogate for ground truth. We take only the low resolution human input that melanocytic atypia exists somewhere, within the huge space of a whole slide image, and use immunohistochemical staining as a labeling mechanism for adjacent H&E sections to train a CNN.

We show our models are able to learn features indicative of melanocyte presence, within H&E images, and perform well on holdout patient samples. Models are robust to similar looking morphologies that do not contain melanocytic atypia, and can utilize information from different stains in alternative but complimentary ways. We are able to map our predictions back onto tissue samples to provide an interpretable view for pathologists to consider, and we observe what morphological features trained networks are activated by.

Creating a working model that is able to adequately identify melanocytic atypia is the first step towards the more complicated task of automating diagnosis of Melanoma, the more deadly form of the skin disease. These models have been successful in a wide range of other projects, and have been a joy to create; it is my aim to publish this work and the work that was not shown here in a manuscript within the next coming months.

2.6 Figures

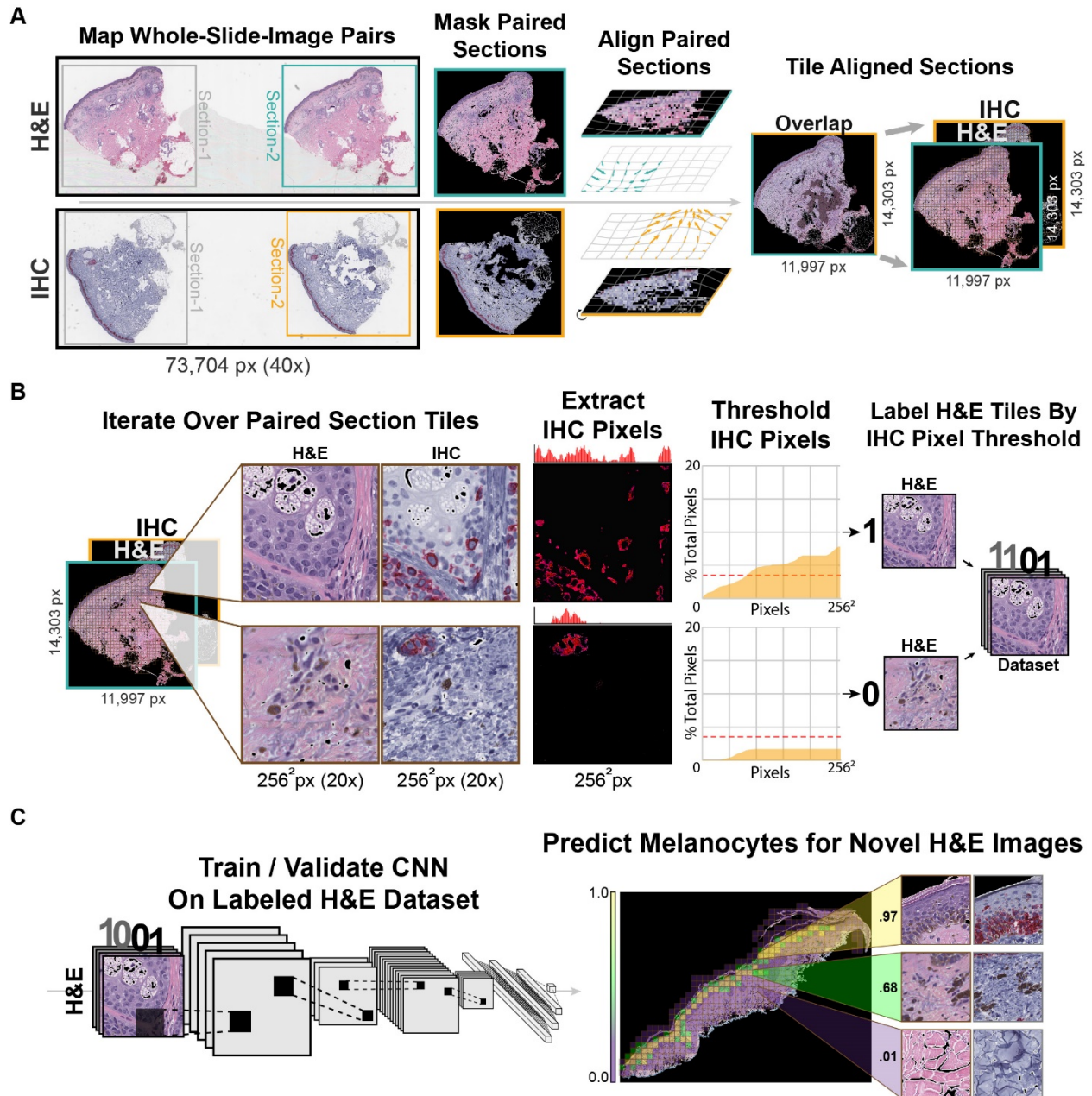


Figure 2.1 - Image processing pipeline. **A)** Matching tissue sections from separate H&E and IHC-stained whole-slide images are extracted and masked, forming a pair. Tissue sections are aligned to each other using Image-J's b-spline alignment algorithm (*BunwarpJ*) in order to ensure optimal congruence at the tile-level. Following alignment, tissue sections are independently tiled. **B)** Labels for H&E tiles are determined by thresholding a color-mask of the corresponding tile's IHC stain. **C)** Training and validation is conducted on only H&E images, with loss calculated by the difference between model output and H&E label, as determined in section B. Once the model is fully trained, predictions for individual tiles from holdout patient tissue form a heatmap.

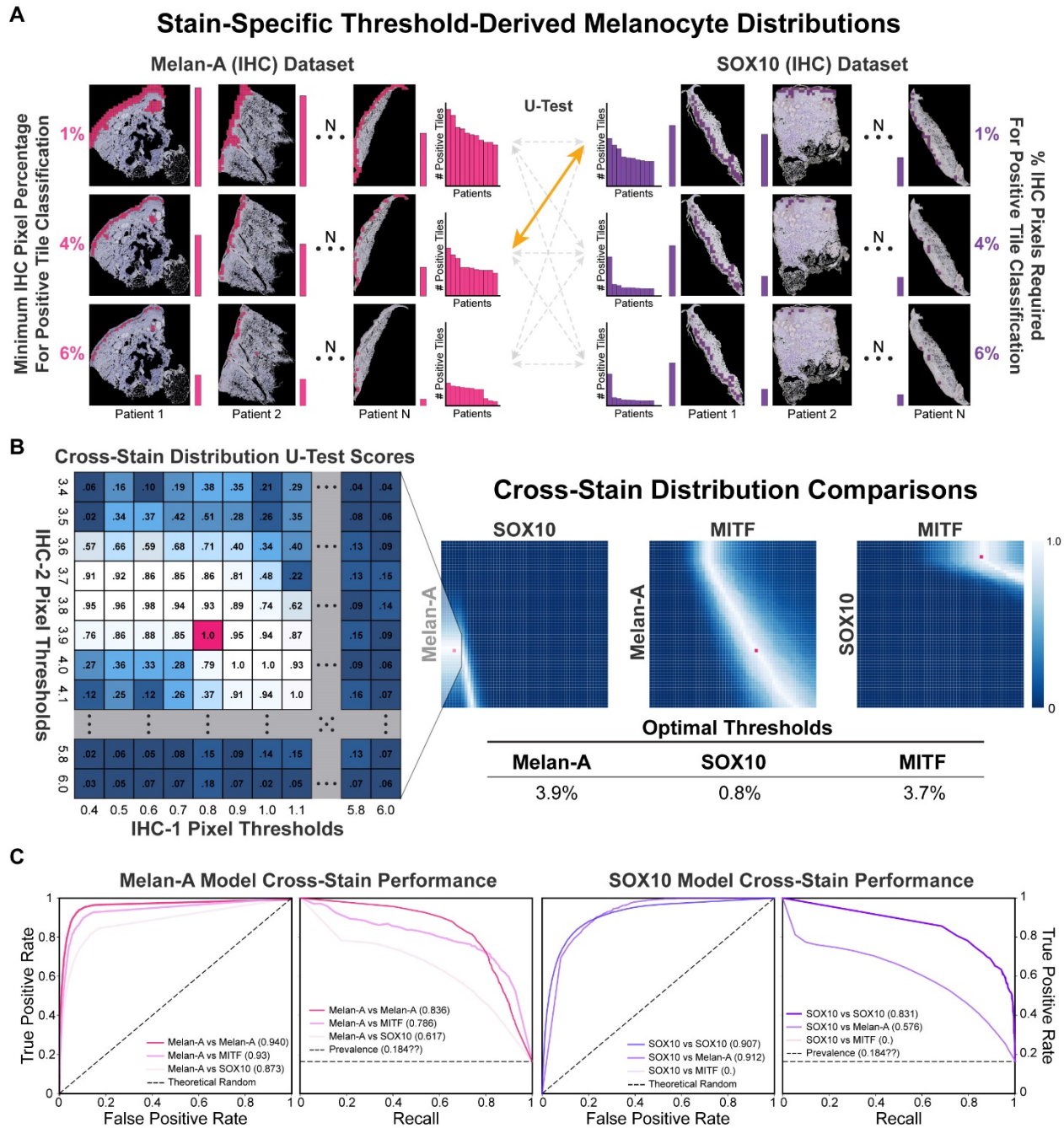


Figure 2.2 - Stain threshold calibration for cross-stain performance metrics. A) A pixel threshold based on the extracted IHC stain is used to label tiles as either likely to contain melanocytes (positive) or unlikely (negative). When comparing across tissue samples stained with different antibodies, the pixel thresholds for each stain should be calibrated so that the average number of melanocytes, across stains, fit a similar distribution. To calibrate pixel thresholds, for each stain, the d for all tissues related to a particular stain were calculated, and a Mann-Whitney U-test was performed to determine distribution similarity. B) Similarity values for distributions, across a range of thresholds. C) Performance metrics for cross-stain predictions, after choosing the thresholds that were most similar across each coupling of stains (red squares, B).

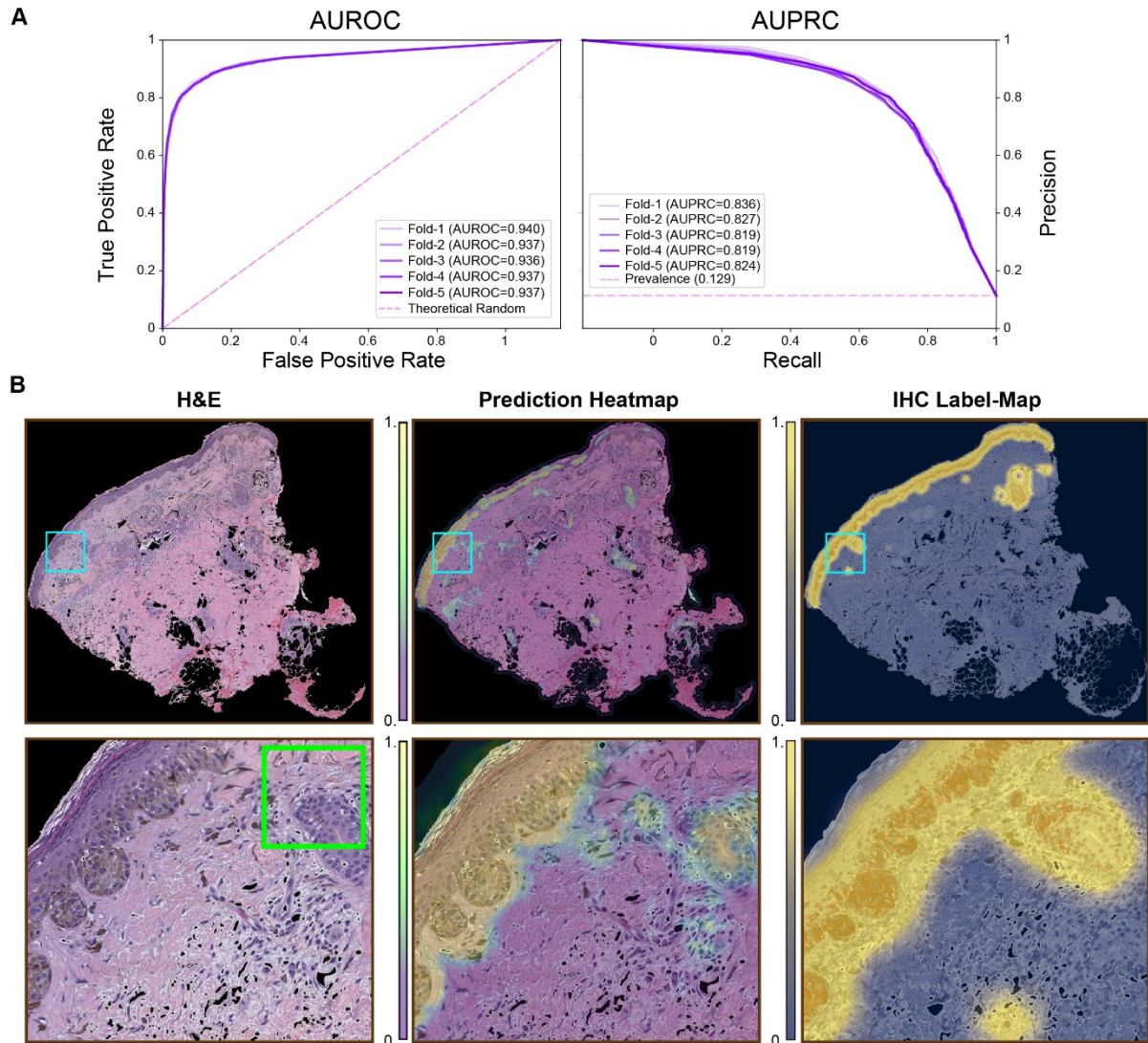


Figure 2.3 – Melan-A Melanocyte-detector performance. **A)** Performance metrics for the model trained on H&E images labeled using a Melan-A or MelPro IHC stain. Performance metrics are evaluated on a holdout patient dataset. The AUROC (left) and AUPRC (right) curves for all runs from a 5-fold cross-validation split are shown. **B)** Melanocyte prediction heatmap. Top panels show the original H&E image, resulting prediction heatmap, and corresponding ground-truth IHC label-map, respectively. Bottom panels show the inset in blue, representing a 3x3 tile region, at higher resolution. The green square represents the size of a single 256x256 tile.

Truth Table Predictions

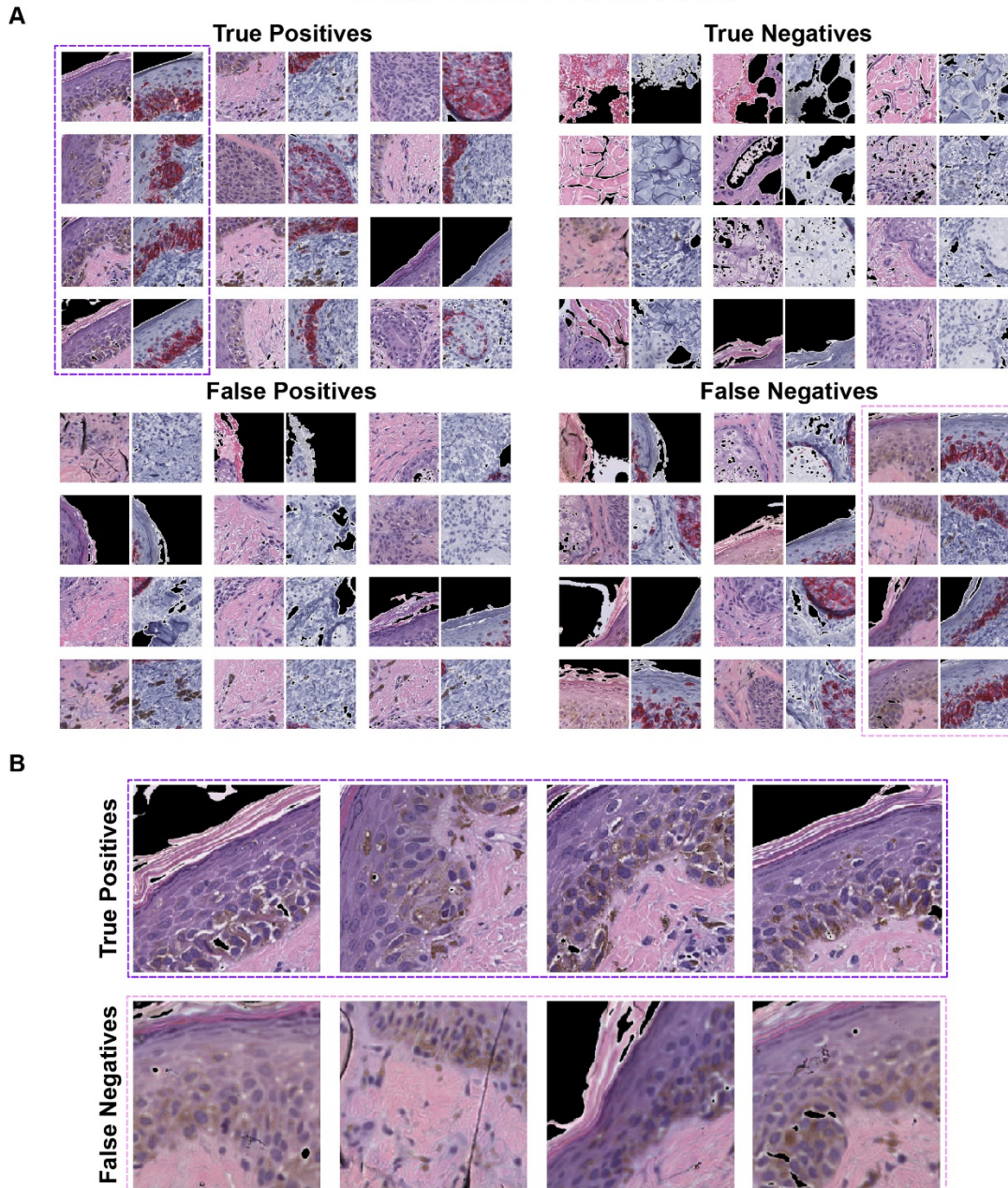
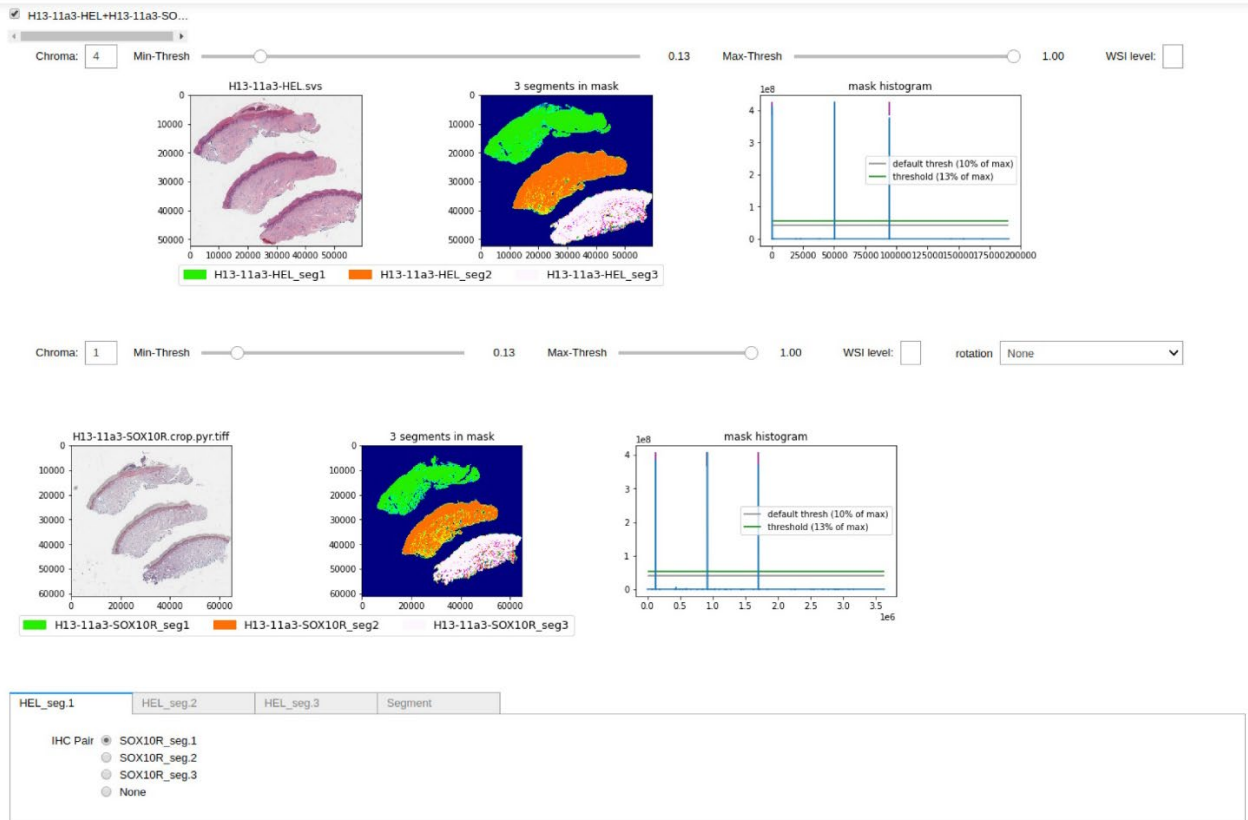
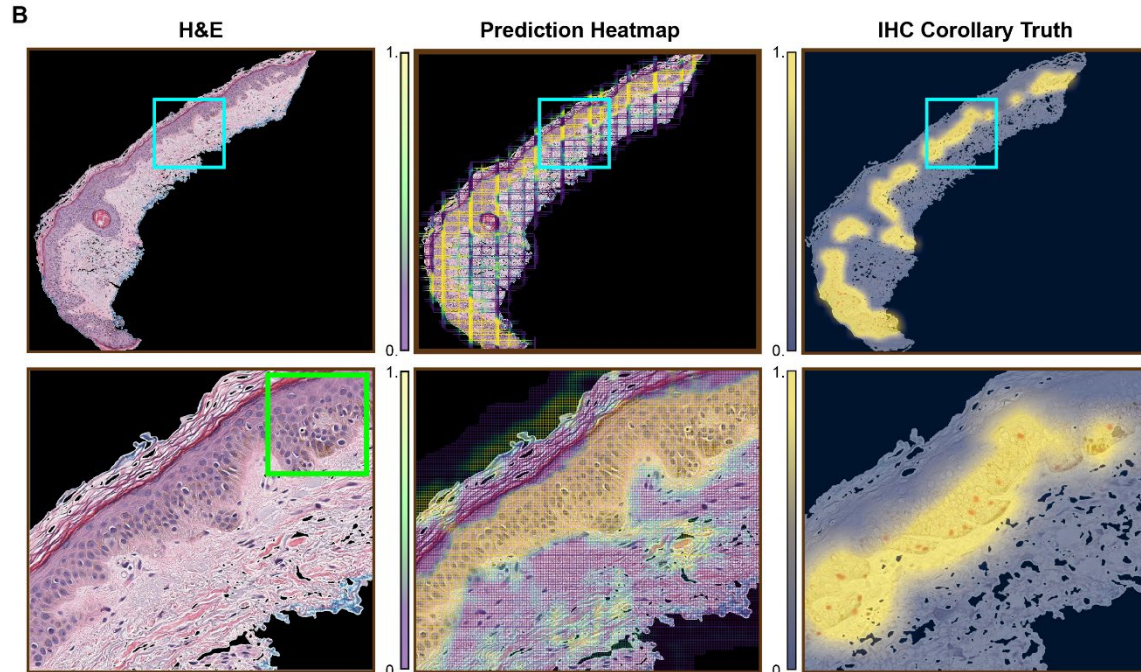
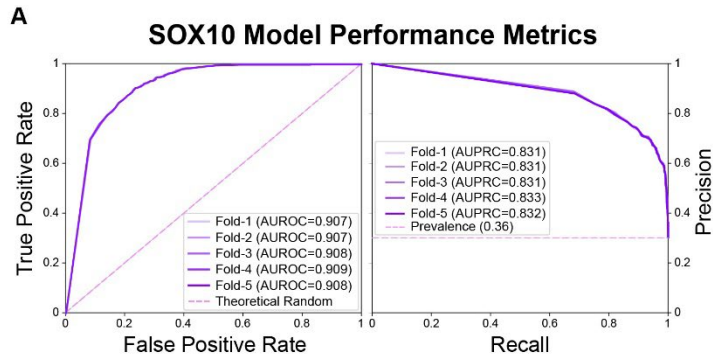


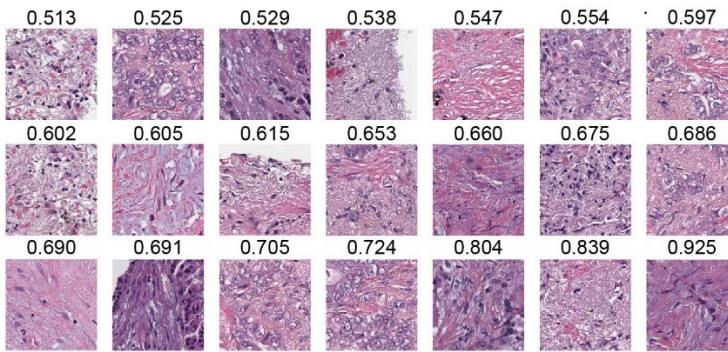
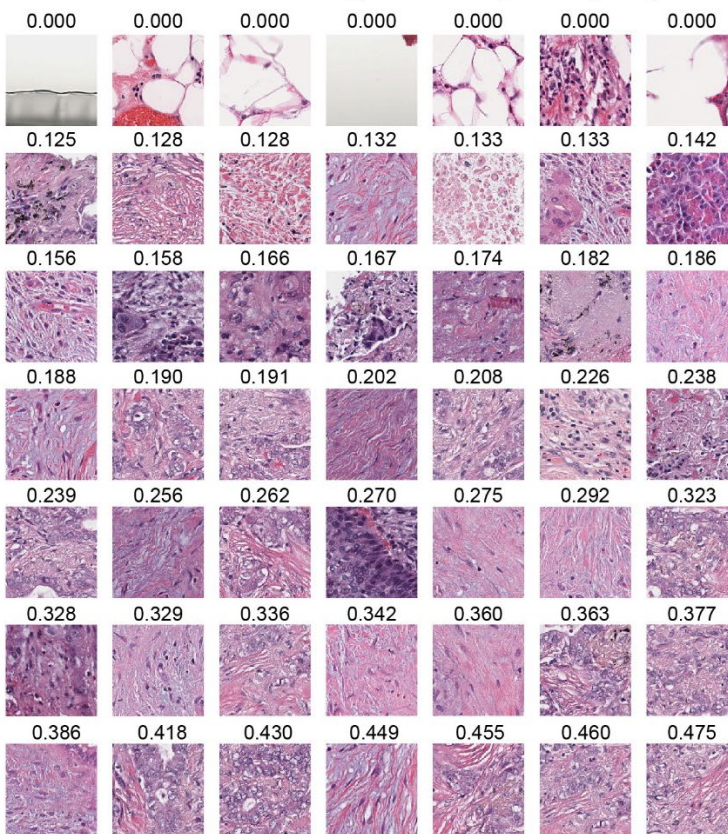
Figure 2.4 - Truth Table Prediction Examples. **A)** Predictions for each tile from a single-patient holdout dataset consisting of ~26000 tiles were assigned to a truth table category. Category membership is based on a 50 percent threshold for positive/negative assignment, and agreement with the ground-truth IHC color-threshold for True/False assignment. Twelve examples were randomly selected from each category. Each example shows the predicted H&E tile as well as its unseen IHC pair, for visual comparison. **B)** Scanning Defaults. While examples from each category provide insight into the model's decision making, a portion of decisions are the (unavoidable) result of damaged tissue or corrupted data during the scanning process. Examples with similar tissue morphologies and localizations highlight the impact blurred regions have on category assignment.



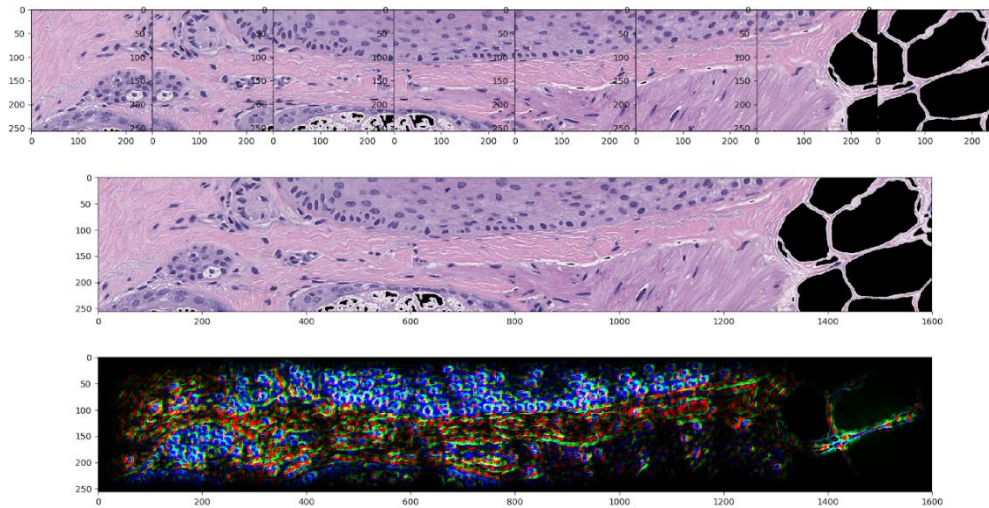
Supplementary Figure A.2.1 – Automated Tissue Extraction Toolkit. An example of a sample pair extracted using our automated toolkit. The toolkit visualizes both the H&E (top, left) and IHC (middle, left) WSIs, and extracts tissue regions according to threshold values set for each of the ipywidgets provided. Matching occurs automatically based on tissue locations in the WSI, and can be manually adjusted by the user if necessary (bottom).



Supplementary Figure A.2.2 – SOX10 Melanocyte-detector performance. **A)** Performance metrics for the model trained on H&E images labeled using a SOX10 IHC stain. Performance metrics are evaluated on a holdout patient dataset. The AUROC (left) and AUPRC (right) curves for all runs from a 5-fold cross-validation split are shown. **B)** Melanocyte prediction heatmap. Top panels show the original H&E image, resulting prediction heatmap, and corresponding ground-truth IHC label-map, respectively. Bottom panels show the inset in blue, representing a 3x3 tile region, at higher resolution. The green square represents the size of a single 256x256 tile.

A**TCGA False Positives (Total)****B****TCGA True Negatives (Sample)**

Supplementary Figure A.2.3 - Truth table predictions from TCGA holdout dataset. To improve generalization of the model, melanocyte-negative tissue samples (pancreas, kidney, and lung) were extracted from TCGA and a random subset of tiles were injected into training and validation datasets. To visualize TCGA tissue diversity and to interrogate model accuracy after training, predictions on a holdout subset of 50,000 randomly selected TCGA tiles were designated as either false-positive or true-negative, based on a prediction threshold of .50. **A)** Of the 50,000 tiles predicted, only 21 tiles (0.00042%) were false-positive retaining a prediction score larger than .50, indicating the models is robust to visually similar features. **B)** Sample true-negative predictions across the negative threshold range are shown.



Supplementary Figure A.2.4 – Saliency Mapping. Saliency mapping highlights, for each tile, which pixels differentially affect the network prediction the most, as indicated by the summed direction of change in their gradients after changes to each pixel value. Individual tiles from a strip of holdout patient tissue are shown on top, as well as a stitched version (middle) where the overlapping regions between tiles have been merged. The bottom image shows the stitched result of applying guided gradients to H&E tiles, and coloring pixels according to the outcome by which they are most activated. Blue/purple regions indicate pixels have more activations when the prediction is positive, while red/green regions are activated more when the prediction is negative.

2.7 Tables

Table 2.1 – SOX10 Whole Slide Image Dataset Overview

Stain Type	Institution	Slice-ID	Sections	Sections	Stain Color	Preview Notes	Action	Alignment
SOX10	UCSF	11-9a	3	2	Black	3 small H&E slices Provided. 2 slightly larger IHC slices -- can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15a1	2	1	Black	2 small slices compared to 1 large rotated slice -- can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15a2	2	1	Black	2 small slices compared to 1 large rotated slice -- can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15a3	4	2	Black	4 slices compared to 2 larger rotated slices only -- can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15a4	2	2	Black	2 small slices compared to 1 large *broken* rotated slice -- can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15a5	2	1	Black	2 small slices compared to 1 large *broken* slice -- can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15a6	2	1	Black	2 small slices compared to 1 large rotated slice -- can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15a7	3	2	Black	Multiple slices compared to 2 separated portion of the same section --can't perform alignment preview	Not Used	Not Aligned
SOX10	UCSF	11-15B	8	4	Black	Tissue is poor/choppy	Not Used	Not Aligned
SOX10	UCSF	12-8	3	3	Red	Good	Used	Aligned
SOX10	UCSF	12-26	8	4	Red	Good -- remove bottom slice tissue broken from IHC -- **WSI throws an error for align upon chroma change	Used	Aligned
SOX10	UCSF	13-6	6	2	Red	Good -- All alignments are likely as good as they will get, but not optimal due to fundamental differences in shape between HE and IHC excisions	Used	Aligned
SOX10	UCSF	13-15	3	3	Red	ffy -- most segments are too differently shaped / sized to incorporate -- May need to use original images rather than ECC	Not Used	Not Aligned
SOX10	UCSF	14-9	12	12	Brown	Good -- STAIN (GREY/BROWN) IS DIFFERENT THAN ORIGINAL LABEL	Used	Aligned
SOX10	UCSF	14-12	6	6	Red	Good -- does not perform well on ECC alignment	Used	Aligned
SOX10	UCSF	15-29	12	4	Red	Poor pairing and can't tel if slice is worthwhile	Not Used	Not Aligned
SOX10	UCSF	15-38	8	8	Red	Good	Attempted	Aligned Poorly
SOX10	UCSF	15-48	6	6	Red	Good -- stain somewhat light	Attempted	Aligned Poorly
SOX10	Stanford	H12-13	3	3	Brown	Spurious -- Slices are huge and IHC stain is hard to separate from BG -- make sure masks are correctly matched	Attempted	Aligned Poorly
SOX10	Stanford	H13-11a3	3	3	Red	COPY of slices in MECLA -- HE images seem like bigger area	Used	Aligned
SOX10	Stanford	H13-11a4	4	3	Red	COPY of slices in MECLA -- Could be hard to align if we don't fix angle disparity	Attempted	Aligned Poorly
SOX10	Stanford	H14-2aa1	4	4	Brown	Good	Attempted	Aligned Poorly
SOX10	Stanford	H14-3b	4	4	Red	Good	Used	Aligned
SOX10	Stanford	H14-7	9	3	Red	Good	Attempted	Aligned Poorly
SOX10	Stanford	H14-15a	1	1	Brown	Good -- See if difference in angle causes alignment to fail	Used	Aligned
SOX10	Stanford	H14-15c	1	1	Brown	Good	Used	Aligned
SOX10	Stanford	H14-17a	4	4	Brown	Good	Used	Aligned
SOX10	Stanford	H15-1a	3	3	Red	Good	Used	Aligned
SOX10	Stanford	H15-7b	3	3	Red	Good - Yellow smudge may cause problem with matches.. make sure they are correct	Used	Aligned

Table 2.2 – Melan-A Whole Slide Image Dataset Overview

Stain Type	Institution	Slice-ID	Sections	Sections	Stain Color	Preview Notes	Action	Alignment
MELA	UCSF	11-3	6	6	Red	Tissue itself is strange. Precarious and previously removed	Not Used	Not Aligned
MELA	UCSF	11-16L1	8	8	Red	3rd level image damaged -- damage may impact alignment. preview 6 slices compared to 2 rotated slices -- 3rd slice on top removed bc of bad alignment	Used	Aligned
MELA	UCSF	12-1a	6	2	Red		Used	Aligned
MELA	UCSF	12-28	2	1	Red	Good -- can't preview due to huge size disparity -- going to assume at highest res it will work out -- **Have to manually add params	Attempted	Aligned Poorly
MELA	UCSF	13-2	3	3	Red	ffy -- 3rd pairing is mismatch at equal sizing so removed -- other two may align, but not likely to very well	Not Used	Not Aligned
MELA	UCSF	13-7	8	8	Red	Good	Used	Aligned
MELA	UCSF	14-11	5	5	Red	Good -- keeping segment 1	Used	Aligned
MELA	UCSF	14-17b	8	4	Red	IHC images on wildly larger scale than HE -- can't preview -- can't ECC align	Not Used	Not Aligned
MELA	UCSF	15-1	8	8	Red	Good -- skip first 2 smallest segments as they are difficult to align	Used	Aligned
MELA	UCSF	15-6	8	8	Red	Good -- remove bottom 2 segments as cannot ECC align between HE and IHC	Attempted	Aligned Poorly
MELA	UCSF	15-17c	4	4	Red	Good -- likely too big for opacity -- skip for now and use previous alignments	Attempted	Aligned Poorly
MELA	UCSF	15-23	4	2	Red	Good	Used	Aligned
MELA	UCSF	15-24a	q	12	Red	Good -- exclude bottom 2 segments by size -- removed section 10-9 HE-IHC for bad alignment	Used	Aligned
MELA	UCSF	15-24b	8	8	Red	Good	Used	Aligned
MELA	UCSF	15-28	18	6	Red	Good -- removed bottom slice from each trio (total of 6 slices) due to bad segmentation	Used	Aligned
MELA	UCSF	15-33a	6	6	Red	May not be able to align due to size disparity	Used	Aligned
MELA	UCSF	15-35	10	10	Red	Good -- ignore middle 2 slices for all pairings	Attempted	Aligned Poorly
MELA	UCSF	15-45	6	2	Red	Poor pairing -- can't preview -- likely can't align	Not Used	Not Aligned
MELA	UCSF	15-46b	5	5	Red	Good -- remove 5<<->5	Attempted	Aligned Poorly
MELA	Stanford	H13-9a	3	3	Brown	Good	Used	Aligned
MELA	Stanford	H13-11a3	3	3	Brown	IHC sections are missing bottom half + angle disparity	Not Used	Not Aligned
MELA	Stanford	H13-11a4	4	4	Brown	IHC sections are missing bottom half + angle disparity	Not Used	Not Aligned
MELA	Stanford	H14-5a	2	2	Brown	HE seg1 contains schlieren lines on right side -- Seg2 shapes different	Attempted	Aligned Poorly
MELA	Stanford	H14-16a2	4	2	Brown	IHC seg2 needs to be resampled	Not Used	Not Aligned
MELA	Stanford	H14-16a3	1	1	Brown	Bad IHC stain	Not Used	Not Aligned
MELA	Stanford	H15-5	1	1	Brown	Good	Attempted	Not Aligned
MelPro	UCSF	15-3	12	4	Red	Good	Not Used	Not Aligned
MelPro	UCSF	15-31	18	6	Red	Good -- bottom 3 IHCs are not attached to slide correctly -- applying only to top 3	Used	Aligned
MelPro	Stanford	H13-7b	6	2	Red	Good -- alignments for pairs 1-3 may be difficult -- Copy in p16	Attempted	Not Aligned
MelPro	Stanford	H13-14c	3	3	Red	Good -- Slight amount of IHC bottom missing in comparison to HEL	Used	Aligned
MelPro	Stanford	H13-23b	3	3	Red	Good -- IHC 2 may be treated as 2 segments instead of 1 (examine whether chroma 0. vs 1. is necessary)	Used	Aligned
MelPro	Stanford	H14-6a	3	3	Red	Good -- NA	Used	Aligned

References

1. Levchenko, K.; Datsenko, O. P.; Serhiichuk, O.; Tolmachev, A.; Iaroshenko, V. O.; Mykhailiuk, P. K. Copper-Catalyzed O-Difluoromethylation of Functionalized Aliphatic Alcohols: Access to Complex Organic Molecules with an Ocf₂h Group. *J. Org. Chem.* **2016**, *81*, 5803-5813.
2. Tolmachev, A.; Bogolubsky, A. V.; Pipko, S. E.; Grishchenko, A. V.; Ushakov, D. V.; Zhemera, A. V.; Viniychuk, O. O.; Konovets, A. I.; Zaporozhets, O. A.; Mykhailiuk, P. K.; Moroz, Y. S. Expanding Synthesizable Space of Disubstituted 1,2,4-Oxadiazoles. *ACS Comb Sci.* **2016**, *18*, 616-624.
3. Bogolubsky, A. V.; Moroz, Y. S.; Mykhailiuk, P. K.; Ostapchuk, E. N.; Rudnichenko, A. V.; Dmytriv, Y. V.; Bondar, A. N.; Zaporozhets, O. A.; Pipko, S. E.; Doroschuk, R. A.; Babichenko, L. N.; Konovets, A. I.; Tolmachev, A. One-Pot Parallel Synthesis of Alkyl Sulfides, Sulfoxides, and Sulfones. *ACS Comb Sci.* **2015**, *17*, 348-354.
4. Bogolubsky, A. V.; Moroz, Y. S.; Mykhailiuk, P. K.; Pipko, S. E.; Zhemera, A. V.; Konovets, A. I.; Stepaniuk, O. O.; Myronchuk, I. S.; Dmytriv, Y. V.; Doroschuk, R. A.; Zaporozhets, O. A.; Tolmachev, A. 2,2,2-Trifluoroethyl Chlorooxoacetate--Universal Reagent for One-Pot Parallel Synthesis of N(1)-Aryl-N(2)-Alkyl-Substituted Oxamides. *ACS Comb Sci.* **2015**, *17*, 615-622.
5. Druzhenko, T.; Denisenko, O.; Kheylik, Y.; Zozulya, S.; Shishkina, S. S.; Tolmachev, A.; Mykhailiuk, P. K. Design, Synthesis, and Characterization of So₂-Containing Azabicyclo[3.N.1]Alkanes: Promising Building Blocks for Drug Discovery. *Org. Lett.* **2015**, *17*, 1922-1925.
6. Bogolubsky, A. V.; Moroz, Y. S.; Mykhailiuk, P. K.; Granat, D. S.; Pipko, S. E.; Konovets, A. I.; Doroschuk, R.; Tolmachev, A. Bis(2,2,2-Trifluoroethyl) Carbonate as a Condensing Agent in One-Pot Parallel Synthesis of Unsymmetrical Aliphatic Ureas. *ACS Comb Sci.* **2014**, *16*, 303-308.

7. Bogolubsky, A. V.; Moroz, Y. S.; Mykhailiuk, P. K.; Panov, D. M.; Pipko, S. E.; Konovets, A. I.; Tolmachev, A. A One-Pot Parallel Reductive Amination of Aldehydes with Heteroaromatic Amines. *ACS Comb Sci.* **2014**, *16*, 375-380.
8. Bogolubsky, A. V.; Moroz, Y. S.; Mykhailiuk, P. K.; Pipko, S. E.; Konovets, A. I.; Sadkova, I. V.; Tolmachev, A. Sulfonyl Fluorides as Alternative to Sulfonyl Chlorides in Parallel Synthesis of Aliphatic Sulfonamides. *ACS Comb Sci.* **2014**, *16*, 192-197.
9. Sterling, T.; Irwin, J. J. Zinc 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324-2337.
10. Teague S. J.; Davis A. M.; Leeson P. D.; Oprea T. I. The Design of Leadlike Combinatorial Libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743-3748.
11. Carr R. A.; Congreve M.; Murray C. W.; Rees D. C. Fragment-Based Lead Discovery: Leads by Design. *Drug Discovery Today.* **2005**, *10*, 987-992.
12. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, 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 Res.* **2014**, *42*, D1083-1090.
13. ZINC Annotated Catalogs. <https://zinc15.docking.org/catalogs/subsets/annotated> (accessed April 10, 2017),
14. Gillet, V. J. New Directions in Library Design and Analysis. *Curr Opin Chem Biol.* **2008**, *12*, 372-378.
15. Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204-3218.
16. Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74-82.

17. Steindl, T. M.; Schuster, D.; Wolber, G.; Laggner, C.; Langer, T. High-Throughput Structure-Based Pharmacophore Modelling as a Basis for Successful Parallel Virtual Screening. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 703-715.
18. Button, A. L.; Hiss, J. A.; Schneider, P.; Schneider, G. Scoring of De Novo Designed Chemical Entities by Macromolecular Target Prediction. *Mol Inform.* **2017**, *36*.
19. Schneider, G.; Schneider, P. Macromolecular Target Prediction by Self-Organizing Feature Maps. *Expert opinion on drug discovery.* **2016**, 1-7.
20. Fu, G.; Nan, X.; Liu, H.; Patel, R. Y.; Daga, P. R.; Chen, Y.; Wilkins, D. E.; Doerksen, R. J. Implementation of Multiple-Instance Learning in Drug Activity Prediction. *BMC Bioinformatics.* **2012**, *13 Suppl 15*, S3.
21. Azencott, C. A.; Ksikes, A.; Swamidass, S. J.; Chen, J. H.; Ralaivola, L.; Baldi, P. One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties. *J. Chem. Inf. Model.* **2007**, *47*, 965-974.
22. Cereto-Massague, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. Tools for in Silico Target Fishing. *Methods.* **2015**, *71*, 98-103.
23. Basak, S. C. Mathematical Descriptors for the Prediction of Property, Bioactivity, and Toxicity of Chemicals from Their Structure: A Chemical-Cum-Biochemical Approach. *Curr Comput Aided Drug Des.* **2013**, *9*, 449-462.
24. Yu, P.; Wild, D. J. Fast Rule-Based Bioactivity Prediction Using Associative Classification Mining. *J Cheminform.* **2012**, *4*, 29.
25. Sugaya, N. Training Based on Ligand Efficiency Improves Prediction of Bioactivities of Ligands and Drug Target Proteins in a Machine Learning Approach. *J. Chem. Inf. Model.* **2013**, *53*, 2525-2537.

26. Murrell, D. S.; Cortes-Ciriano, I.; van Westen, G. J.; Stott, I. P.; Bender, A.; Malliavin, T. E.; Glen, R. C. Chemically Aware Model Builder (Camb): An R Package for Property and Bioactivity Modelling of Small Molecules. *J Cheminform.* **2015**, *7*, 45.
27. Seal, A.; Ahn, Y. Y.; Wild, D. J. Optimizing Drug-Target Interaction Prediction Based on Random Walk on Heterogeneous Networks. *J Cheminform.* **2015**, *7*, 40.
28. Iskar, M.; Zeller, G.; Zhao, X. M.; van Noort, V.; Bork, P. Drug Discovery in the Age of Systems Biology: The Rise of Computational Approaches for Data Integration. *Curr Opin Biotechnol.* **2012**, *23*, 609-616.
29. Peragovics, A.; Simon, Z.; Tombor, L.; Jelinek, B.; Hari, P.; Czobor, P.; Malnasi-Csizmadia, A. Virtual Affinity Fingerprints for Target Fishing: A New Application of Drug Profile Matching. *J. Chem. Inf. Model.* **2013**, *53*, 103-113.
30. Chen, X.; Yan, C. C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. Drug-Target Interaction Prediction: Databases, Web Servers and Computational Models. *Briefings in bioinformatics.* **2016**, *17*, 696-712.
31. Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X. Q. Targethunter: An in Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS J.* **2013**, *15*, 395-406.
32. Chen, X.; Liang, Y.; Xu, J. Toward Automated Biochemotype Annotation for Large Compound Libraries. *Mol Divers.* **2006**, *10*, 495-509.
33. Nguyen, H. P.; Koutsoukas, A.; Mohd Fauzi, F.; Drakakis, G.; Maciejewski, M.; Glen, R. C.; Bender, A. Diversity Selection of Compounds Based on 'Protein Affinity Fingerprints' Improves Sampling of Bioactive Chemical Space. *Chem. Biol. Drug Des.* **2013**, *82*, 252-266.
34. Huang, T.; Mi, H.; Lin, C. Y.; Zhao, L.; Zhong, L. L.; Liu, F. B.; Zhang, G.; Lu, A. P.; Bian, Z. X.; for, M. G. Most: Most-Similar Ligand Based Approach to Target Prediction. *BMC Bioinformatics.* **2017**, *18*, 165.

35. Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. *J. Med. Chem.* **2011**, *54*, 1223-1232.
36. Wolber, G.; Dornhofer, A. A.; Langer, T. Efficient Overlay of Small Organic Molecules Using 3d Pharmacophores. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 773-788.
37. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat Biotechnol.* **2007**, *25*, 197-206.
38. Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijter, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting New Molecular Targets for Known Drugs. *Nature.* **2009**, *462*, 175-181.
39. 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. *J. Med. Chem.* **2010**, *53*, 2464-2471.
40. Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Cote, S.; Shoichet, B. K.; Urban, L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature.* **2012**, *486*, 361-367.
41. Laggner, C.; Kokel, D.; Setola, V.; Tolia, A.; Lin, H.; Irwin, J. J.; Keiser, M. J.; Cheung, C. Y.; Minor, D. L., Jr.; Roth, B. L.; Peterson, R. T.; Shoichet, B. K. Chemical Informatics and Target Identification in a Zebrafish Phenotypic Screen. *Nat Chem Biol.* **2012**, *8*, 144-146.
42. 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 Biol.* **2013**, *11*, e1001712.
43. Tanimoto, T. T. *IBM Internal Report.* **1957**.

44. Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941-948.
45. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742-754.
46. Koutsoukas, A.; Lowe, R.; Kalantarmotamedi, Y.; Mussa, H. Y.; Klaffke, W.; Mitchell, J. B.; Glen, R. C.; Bender, A. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naive Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* **2013**, *53*, 1957-1966.
47. Gregori-Puigjane, E.; Setola, V.; Hert, J.; Crews, B. A.; Irwin, J. J.; Lounkine, E.; Marnett, L.; Roth, B. L.; Shoichet, B. K. Identifying Mechanism-of-Action Targets for Drugs and Probes. *Proc Natl Acad Sci U S A.* **2012**, *109*, 11178-11183.
48. Irwin J. J.; Duan D.; Torosyan H.; Doak A. K.; Ziebart K. T.; Sterling T.; Tumanian G.; Shoichet B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076–7087.
49. Irwin J. J.; Shoichet B. K. Docking Screens for Novel Ligands Conferring New Biology. *J. Med. Chem.* **2016**, *59*, 4103–4120
50. Aldrich C.; Bertozzi C.; Georg G. I.; Kiessling L.; Lindsley C.; Liotta D.; Merz K. M. Jr.; Schepartz A.; Wang S. The Ecstasy and Agony of Assay Interference Compounds. *J. Med. Chem.* **2017**, *60*, 2165–2168.
51. ZINC Genes Having Ligands of 10um or Better, after the Treatment for Pains and Aggregator Analogs Described in the Methods. https://zinc15.docking.org/genes/?num_substances-gt=0&num_predictions-gt=0 (accessed May 27, 2017),
52. Inui, J.; Imamura, H. Beta-Adrenoceptor Blocking and Electrophysiological Effects of Bufetolol in the Guinea Pig Atria. *Eur J Pharmacol.* **1977**, *41*, 251-260.

53. Masumiya, H.; Tanaka, Y.; Tanaka, H.; Shigenobu, K. Inhibition of T-Type and L-Type Ca(2+) Currents by Aranidipine, a Novel Dihydropyridine Ca(2+) Antagonist. *Pharmacology*. **2000**, *61*, 57-61.
54. Nelson, K. M.; Dahlin, J. L.; Bisson, J.; Graham, J.; Pauli, G. F.; Walters, M. A. The Essential Medicinal Chemistry of Curcumin. *J. Med. Chem.* **2017**, *60*, 1620-1637.
55. Saario, S. M.; Poso, A.; Juvonen, R. O.; Jarvinen, T.; Salo-Ahen, O. M. Fatty Acid Amide Hydrolase Inhibitors from Virtual Screening of the Endocannabinoid System. *J. Med. Chem.* **2006**, *49*, 4650-4656.
56. Bray, N. Lead Identification: Shedding Light on Dark Chemical Matter. *Nat Rev Drug Discov.* **2015**, *14*, 817.
57. Macarron, R. Chemical Libraries: How Dark Is Hts Dark Matter? *Nat Chem Biol.* **2015**, *11*, 904-905.
58. Muegge, I.; Mukherjee, P. Performance of Dark Chemical Matter in High Throughput Screening. *J. Med. Chem.* **2016**, *59*, 9806-9813.
59. Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat Chem Biol.* **2015**, *11*, 958-966.
60. ZINC Zinc15 Resources Wiki Page. <http://wiki.docking.org/index.php/ZINC15:Resources> (accessed Oct 12, 2015),
61. ZINC Genes Having 20 or More Ligands Where None Is for Sale, and Predictions. https://zinc15.docking.org/genes/?num_purchasable=0&num_predictions-gt=0&num_substances-gt=20 (accessed May 27, 2017),
62. Hert, J.; Irwin, J. J.; Laggner, C.; Keiser, M. J.; Shoichet, B. K. Quantifying Biogenic Bias in Screening Libraries. *Nat Chem Biol.* **2009**, *5*, 479-483.

63. Chavez-Pina, A. E.; McKnight, W.; Dickey, M.; Castaneda-Hernandez, G.; Wallace, J. L. Mechanisms Underlying the Anti-Inflammatory Activity and Gastric Safety of Acemetacin. *Br J Pharmacol.* **2007**, *152*, 930-938.
64. Mross, K.; Richly, H.; Schleucher, N.; Korfee, S.; Tewes, M.; Scheulen, M. E.; Seeber, S.; Beinert, T.; Schweigert, M.; Sauer, U.; Unger, C.; Behringer, D.; Brendel, E.; Haase, C. G.; Voliotis, D.; Strumberg, D. A Phase I Clinical and Pharmacokinetic Study of the Camptothecin Glycoconjugate, Bay 38-3441, as a Daily Infusion in Patients with Advanced Solid Tumors. *Ann Oncol.* **2004**, *15*, 1284-1294.
65. Hofmann, T. G.; Hehner, S. P.; Bacher, S.; Droge, W.; Schmitz, M. L. Various Glucocorticoids Differ in Their Ability to Induce Gene Expression, Apoptosis and to Repress Nf-KappaB-Dependent Transcription. *FEBS Lett.* **1998**, *441*, 441-446.
66. Raguene-Nicol, C.; Russo-Marie, F.; Domage, G.; Diab, N.; Solito, E.; Dray, F.; Mace, J. L.; Streichenberger, G. Anti-Inflammatory Mechanism of Alminoprofen: Action on the Phospholipid Metabolism Pathway. *Biochem. Pharmacol.* **1999**, *57*, 433-443.
67. Schoemaker, H.; Claustre, Y.; Fage, D.; Rouquier, L.; Chergui, K.; Curet, O.; Oblin, A.; Gonon, F.; Carter, C.; Benavides, J.; Scatton, B. Neurochemical Characteristics of Amisulpride, an Atypical Dopamine D2/D3 Receptor Antagonist with Both Presynaptic and Limbic Selectivity. *J Pharmacol Exp Ther.* **1997**, *280*, 83-97.
68. McLean, R. C.; Baird, S. W.; Becker, L. C.; Townsend, S. N.; Gerstenblith, G.; Kass, D. A.; Tomaselli, G. F.; Schulman, S. P. Response to Catecholamine Stimulation of Polymorphisms of the Beta-1 and Beta-2 Adrenergic Receptors. *Am J Cardiol.* **2012**, *110*, 1001-1007.
69. Tsukagoshi, S. [Pharmacokinetics of Azasetron (Serotone), a Selective 5-Ht3 Receptor Antagonist]. *Gan To Kagaku Ryoho.* **1999**, *26*, 1001-1008.
70. Oizumi, K.; Nishino, H.; Koike, H.; Sada, T.; Miyamoto, M.; Kimura, T. Antihypertensive Effects of Cs-905, a Novel Dihydropyridine Ca⁺⁺ Channel Blocker. *Jpn. J. Pharmacol.* **1989**, *51*, 57-64.

71. Yamamoto, M.; Shimizu, M. Effects of a New Trh Analogue, Ym-14673 on the Central Nervous System. *Naunyn Schmiedebergs Arch Pharmacol.* **1987**, *336*, 561-565.
72. Cambau, E.; Matrat, S.; Pan, X. S.; Roth Dit Bettoni, R.; Corbel, C.; Aubry, A.; Lascols, C.; Driot, J. Y.; Fisher, L. M. Target Specificity of the New Fluoroquinolone Besifloxacin in Streptococcus Pneumoniae, Staphylococcus Aureus and Escherichia Coli. *J. Antimicrob. Chemother.* **2009**, *63*, 443-450.
73. Frishman, W. H.; Goldberg, R. J.; Benfield, P. Bevantolol. A Preliminary Review of Its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy in Hypertension and Angina Pectoris. *Drugs.* **1988**, *35*, 1-21.
74. Corcostegui, R.; Labeaga, L.; Innerarity, A.; Berisa, A.; Orjales, A. Preclinical Pharmacology of Bilastine, a New Selective Histamine H1 Receptor Antagonist: Receptor Selectivity and in Vitro Antihistaminic Activity. *Drugs R D.* **2005**, *6*, 371-384.
75. Bertrand, F.; Lehmann, O.; Galani, R.; Lazarus, C.; Jeltsch, H.; Cassel, J. C. Effects of Mdl 73005 on Water-Maze Performances and Locomotor Activity in Scopolamine-Treated Rats. *Pharmacol Biochem Behav.* **2001**, *68*, 647-660.
76. Hara, H.; Ichikawa, M.; Oku, H.; Shimazawa, M.; Araie, M. Bunazosin, a Selective Alpha1-Adrenoceptor Antagonist, as an Anti-Glaucoma Drug: Effects on Ocular Circulation and Retinal Neuronal Damage. *Cardiovasc Drug Rev.* **2005**, *23*, 43-56.
77. Malinowska, B.; Kiec-Kononowicz, K.; Flau, K.; Godlewski, G.; Kozłowska, H.; Kathmann, M.; Schlicker, E. Atypical Cardiostimulant Beta-Adrenoceptor in the Rat Heart: Stereoselective Antagonism by Bupranolol but Lack of Effect by Some Bupranolol Analogues. *Br J Pharmacol.* **2003**, *139*, 1548-1554.
78. Houin, G.; Barre, J.; Jeannot, J. P.; Ledudal, P.; Cautreels, W.; Tillement, J. P. Pharmacokinetics of Butofilolol (Cafide) after Repeated Oral Administration in Man. *Int J Clin Pharmacol Res.* **1984**, *4*, 175-183.

79. Holick, M. F.; DeLuca, H. F.; Avioli, L. V. Isolation and Identification of 25-Hydroxycholecalciferol from Human Plasma. *Arch Intern Med.* **1972**, *129*, 56-61.
80. Shibuya, T.; Field, R.; Watanabe, Y.; Sato, K.; Salafsky, B. Structure-Affinity Relationships between Several New Benzodiazepine Derivatives and 3h-Diazepam Receptor Sites. *Jpn. J. Pharmacol.* **1984**, *34*, 435-440.
81. Fulton, B.; Markham, A. Mycophenolate Mofetil. A Review of Its Pharmacodynamic and Pharmacokinetic Properties and Clinical Efficacy in Renal Transplantation. *Drugs.* **1996**, *51*, 278-298.
82. Fischli, W.; Clozel, J. P.; Breu, V.; Buchmann, S.; Mathews, S.; Stadler, H.; Vieira, E.; Wostl, W. Ciprokiren (Ro 44-9375). A Renin Inhibitor with Increasing Effects on Chronic Treatment. *Hypertension.* **1994**, *24*, 163-169.
83. Chen, Z.; Skolnick, P. Triple Uptake Inhibitors: Therapeutic Potential in Depression and Beyond. *Expert Opin Investig Drugs.* **2007**, *16*, 1365-1377.
84. Ward, D. A.; Abney, K.; Oliver, J. W. The Effects of Topical Ocular Application of 0.25% Demecarium Bromide on Serum Acetylcholinesterase Levels in Normal Dogs. *Vet Ophthalmol.* **2003**, *6*, 23-25.
85. Gorrill, M. J.; Marshall, J. R. Pharmacology of Estrogens and Estrogen-Induced Effects on Nonreproductive Organs and Systems. *J Reprod Med.* **1986**, *31*, 842-847.
86. Dietz, M.; Mohr, P.; Kuhn, B.; Maerki, H. P.; Hartman, P.; Ruf, A.; Benz, J.; Grether, U.; Wright, M. B. Comparative Molecular Profiling of the Pparalpha/Gamma Activator Aleglitazar: Ppar Selectivity, Activity and Interaction with Cofactors. *ChemMedChem.* **2012**, *7*, 1101-1111.
87. Tanaka, H.; Shigenobu, K. Efonidipine Hydrochloride: A Dual Blocker of L- and T-Type Ca(2+) Channels. *Cardiovasc Drug Rev.* **2002**, *20*, 81-92.
88. Nabeshima, T.; Matsuno, K.; Kamei, H.; Kameyama, T. The Interaction of Eptazocine, a Novel Analgesic, with Opioid Receptors. *Res. Commun. Chem. Pathol. Pharmacol.* **1985**, *48*, 173-181.

89. Zarbin, M. A.; Palacios, J. M.; Wamsley, J. K.; Kuhar, M. J. Axonal Transport of Beta-Adrenergic Receptors. Antero- and Retrogradely Transported Receptors Differ in Agonist Affinity and Nucleotide Sensitivity. *Mol Pharmacol.* **1983**, *24*, 341-348.
90. Aasmundstad, T. A.; Xu, B. Q.; Johansson, I.; Ripel, A.; Bjorneboe, A.; Christophersen, A. S.; Bodd, E.; Morland, J. Biotransformation and Pharmacokinetics of Ethylmorphine after a Single Oral Dose. *Br. J. Clin. Pharmacol.* **1995**, *39*, 611-620.
91. Schlaepfer, I. R.; Rider, L.; Rodrigues, L. U.; Gijon, M. A.; Pac, C. T.; Romero, L.; Cimic, A.; Sirintrapun, S. J.; Glode, L. M.; Eckel, R. H.; Cramer, S. D. Lipid Catabolism Via Cpt1 as a Therapeutic Target for Prostate Cancer. *Mol. Cancer Ther.* **2014**, *13*, 2361-2371.
92. Brune, M. E.; Katwala, S. P.; Milicic, I.; Witte, D. G.; Kerwin, J. F., Jr.; Meyer, M. D.; Hancock, A. A.; Williams, M. Effect of Fiduxosin, an Antagonist Selective for Alpha(1a)- and Alpha(1d)-Adrenoceptors, on Intraurethral and Arterial Pressure Responses in Conscious Dogs. *J Pharmacol Exp Ther.* **2002**, *300*, 487-494.
93. Sutherland, R.; Croydon, E. A.; Rolinson, G. N. Flucloxacillin, a New Isoxazolyl Penicillin, Compared with Oxacillin, Cloxacillin, and Dicloxacillin. *Br Med J.* **1970**, *4*, 455-460.
94. Brodzki, M.; Rutkowski, R.; Jateczak, M.; Kisiel, M.; Czyzewska, M. M.; Mozrzymas, J. W. Comparison of Kinetic and Pharmacological Profiles of Recombinant Alpha1gamma2l and Alpha1beta2gamma2l Gaba_a Receptors - a Clue to the Role of Intersubunit Interactions. *Eur J Pharmacol.* **2016**, *784*, 81-89.
95. Louca Jounger, S.; Christidis, N.; Hedenberg-Magnusson, B.; List, T.; Svensson, P.; Schalling, M.; Ernberg, M. Influence of Polymorphisms in the Htr3a and Htr3b Genes on Experimental Pain and the Effect of the 5-Ht₃ Antagonist Granisetron. *PLoS One.* **2016**, *11*, e0168703.
96. Halobetasol Propionate: A Trihalogenated Ultrapotent Topical Corticosteroid. *J Am Acad Dermatol.* **1991**, *25*, 1137-1186.

97. Pinder, R. M.; Brogden, R. N.; Speight, T. M.; Avery, G. S. Hexoprenaline: A Review of Its Pharmacological Properties and Therapeutic Efficacy with Particular Reference to Asthma. *Drugs*. **1977**, *14*, 1-28.
98. Christensen, C. B. The Opioid Receptor Binding Profiles of Ketobemidone and Morphine. *Pharmacol Toxicol*. **1993**, *73*, 344-345.
99. Klotz, U. Interaction Potential of Lercanidipine, a New Vasoselective Dihydropyridine Calcium Antagonist. *Arzneimittelforschung*. **2002**, *52*, 155-161.
100. Dilworth, F. J.; Williams, G. R.; Kissmeyer, A. M.; Nielsen, J. L.; Binderup, E.; Calverley, M. J.; Makin, H. L.; Jones, G. The Vitamin D Analog, Kh1060, Is Rapidly Degraded Both in Vivo and in Vitro Via Several Pathways: Principal Metabolites Generated Retain Significant Biological Activity. *Endocrinology*. **1997**, *138*, 5485-5496.
101. Holmes, B.; Ward, A. Meptazinol. A Review of Its Pharmacodynamic and Pharmacokinetic Properties and Therapeutic Efficacy. *Drugs*. **1985**, *30*, 285-312.
102. Brooks, A. M.; Gillies, W. E. Ocular Beta-Blockers in Glaucoma Management. Clinical Pharmacological Aspects. *Drugs Aging*. **1992**, *2*, 208-221.
103. Shelly, W.; Draper, M. W.; Krishnan, V.; Wong, M.; Jaffe, R. B. Selective Estrogen Receptor Modulators: An Update on Recent Clinical Findings. *Obstet Gynecol Surv*. **2008**, *63*, 163-181.
104. Gustavo, R. P. [Anti-Gonadotropic Action of Possipione]. *Quad Clin Ostet Ginecol*. **1958**, *13*, 307-315.
105. Attwood, D. Aggregation of Antiacetylcholine Drugs in Aqueous Solution: Monomer Concentrations in Non-Micellar Drug Systems. *J. Pharm. Pharmacol*. **1976**, *28*, 762-765.
106. Marks, M. J.; Wageman, C. R.; Grady, S. R.; Gopalakrishnan, M.; Briggs, C. A. Selectivity of Abt-089 for Alpha4beta2* and Alpha6beta2* Nicotinic Acetylcholine Receptors in Brain. *Biochem. Pharmacol*. **2009**, *78*, 795-802.

107. Muraki, Y. Comparative Functional Selectivity of Imidafenacin and Propiverine, Antimuscarinic Agents, for the Urinary Bladder over Colon in Conscious Rats. *Naunyn Schmiedebergs Arch Pharmacol.* **2015**, *388*, 1171-1178.
108. McGorum, B. C.; Nicholas, D. R.; Foster, A. P.; Shaw, D. J.; Pirie, R. S. Bronchodilator Activity of the Selective Muscarinic Antagonist Revatropate in Horses with Heaves. *Vet J.* **2013**, *195*, 80-85.
109. Greenblatt, D. J. Pharmacology of Benzodiazepine Hypnotics. *J Clin Psychiatry.* **1992**, *53 Suppl*, 7-13.
110. Kang, S. G.; Kim, J. J. Udenafil: Efficacy and Tolerability in the Management of Erectile Dysfunction. *Ther Adv Urol.* **2013**, *5*, 101-110.
111. Abe, S.; Watabe, H.; Takaseki, S.; Aihara, M.; Yoshitomi, T. The Effects of Prostaglandin Analogues on Intracellular Ca²⁺ in Ciliary Arteries of Wild-Type and Prostanoid Receptor-Deficient Mice. *J. Ocul. Pharmacol. Ther.* **2013**, *29*, 55-60.
112. Halland, N.; Blum, H.; Buning, C.; Kohlmann, M.; Lindenschmidt, A. Small Macrocycles as Highly Active Integrin Alpha2beta1 Antagonists. *ACS Med Chem Lett.* **2014**, *5*, 193-198.
113. Lopus, M.; Smiyun, G.; Miller, H.; Oroudjev, E.; Wilson, L.; Jordan, M. A. Mechanism of Action of Ixabepilone and Its Interactions with the Betaiii-Tubulin Isotype. *Cancer Chemother Pharmacol.* **2015**, *76*, 1013-1024.
114. Myatt, J. W.; Healy, M. P.; Bravi, G. S.; Billinton, A.; Johnson, C. N.; Matthews, K. L.; Jandu, K. S.; Meng, W.; Hersey, A.; Livermore, D. G.; Douault, C. B.; Witherington, J.; Bit, R. A.; Rowedder, J. E.; Brown, J. D.; Clayton, N. M. Pyrazolopyridazine Alpha-2-Delta-1 Ligands for the Treatment of Neuropathic Pain. *Bioorg Med Chem Lett.* **2010**, *20*, 4683-4688.
115. van Hattum, H.; Branderhorst, H. M.; Moret, E. E.; Nilsson, U. J.; Leffler, H.; Pieters, R. J. Tuning the Preference of Thiodigalactoside- and Lactosamine-Based Ligands to Galectin-3 over Galectin-1. *J. Med. Chem.* **2013**, *56*, 1350-1354.

116. Saku, O.; Ohta, K.; Arai, E.; Nomoto, Y.; Miura, H.; Nakamura, H.; Fuse, E.; Nakasato, Y.
Synthetic Study of V α 4/Vcam-1 Inhibitors: Synthesis and Structure-Activity Relationship of
Piperazinylphenylalanine Derivatives. *Bioorg Med Chem Lett.* **2008**, *18*, 1053-1057.
117. Ulven, T.; Frimurer, T. M.; Receveur, J. M.; Little, P. B.; Rist, O.; Norregaard, P. K.; Hogberg,
T. 6-Acylamino-2-Aminoquinolines as Potent Melanin-Concentrating Hormone 1 Receptor
Antagonists. Identification, Structure-Activity Relationship, and Investigation of Binding Mode. *J.*
Med. Chem. **2005**, *48*, 5684-5697.
118. Scott, J. S.; Berry, D. J.; Brown, H. S.; Buckett, L.; Clarke, D. S.; Goldberg, K.; Hudson, J. A.;
Leach, A. G.; MacFaul, P. A.; Raubo, P.; Robb, G. Achieving Improved Permeability by
Hydrogen Bond Donor Modulation in a Series of Mgat2 Inhibitors. *MedChemComm.* **2013**, *4*,
1305-1311.
119. Whitby, R. J.; Stec, J.; Blind, R. D.; Dixon, S.; Leesnitzer, L. M.; Orband-Miller, L. A.;
Williams, S. P.; Willson, T. M.; Xu, R.; Zuercher, W. J.; Cai, F.; Ingraham, H. A. Small Molecule
Agonists of the Orphan Nuclear Receptors Steroidogenic Factor-1 (Sf-1, Nr5a1) and Liver
Receptor Homologue-1 (Lrh-1, Nr5a2). *J. Med. Chem.* **2011**, *54*, 2266-2281.
120. DeNinno, M. P.; Wright, S. W.; Etienne, J. B.; Olson, T. V.; Rocke, B. N.; Corbett, J. W.; Kung,
D. W.; DiRico, K. J.; Andrews, K. M.; Millham, M. L.; Parker, J. C.; Esler, W.; van Volkenburg,
M.; Boyer, D. D.; Houseknecht, K. L.; Doran, S. D. Discovery of Triazolopyrimidine-Based
Pde8b Inhibitors: Exceptionally Ligand-Efficient and Lipophilic Ligand-Efficient Compounds for
the Treatment of Diabetes. *Bioorg Med Chem Lett.* **2012**, *22*, 5721-5726.
121. Preuss, J.; Maloney, P.; Peddibhotla, S.; Hedrick, M. P.; Hershberger, P.; Gosalia, P.; Milewski,
M.; Li, Y. L.; Sugarman, E.; Hood, B.; Suyama, E.; Nguyen, K.; Vasile, S.; Sergienko, E.;
Mangravita-Novo, A.; Vicchiarelli, M.; McAnally, D.; Smith, L. H.; Roth, G. P.; Diwan, J.;
Chung, T. D.; Jortzik, E.; Rahlfs, S.; Becker, K.; Pinkerton, A. B.; Bode, L. Discovery of a
Plasmodium Falciparum Glucose-6-Phosphate Dehydrogenase 6-Phosphogluconolactonase

- Inhibitor (R,Z)-N-((1-Ethylpyrrolidin-2-Yl)Methyl)-2-(2-Fluorobenzylidene)-3-Oxo-3,4-Dihydro-2H-Benzo[B][1,4]Thiazine-6-Carboxamide (M1276) That Reduces Parasite Growth in Vitro. *J. Med. Chem.* **2012**, *55*, 7262-7272.
122. Faull, A. W.; Brewster, A. G.; Brown, G. R.; Smithers, M. J.; Jackson, R. Dual-Acting Thromboxane Receptor Antagonist/Synthase Inhibitors: Synthesis and Biological Properties of [2-Substituted-4-(3-Pyridyl)-1,3-Dioxan-5-Yl] Alkenoic Acids. *J. Med. Chem.* **1995**, *38*, 686-694.
123. Kumar, N. S.; Amandoron, E. A.; Cherkasov, A.; Finlay, B. B.; Gong, H.; Jackson, L.; Kaur, S.; Lian, T.; Moreau, A.; Labriere, C.; Reiner, N. E.; See, R. H.; Strynadka, N. C.; Thorson, L.; Wong, E. W.; Worrall, L.; Zoraghi, R.; Young, R. N. Optimization and Structure-Activity Relationships of a Series of Potent Inhibitors of Methicillin-Resistant Staphylococcus Aureus (Mrsa) Pyruvate Kinase as Novel Antimicrobial Agents. *Bioorg Med Chem.* **2012**, *20*, 7069-7082.
124. Drouin, L.; McGrath, S.; Vidler, L. R.; Chaikuad, A.; Monteiro, O.; Tallant, C.; Philpott, M.; Rogers, C.; Fedorov, O.; Liu, M.; Akhtar, W.; Hayes, A.; Raynaud, F.; Muller, S.; Knapp, S.; Hoelder, S. Structure Enabled Design of Baz2-Icr, a Chemical Probe Targeting the Bromodomains of Baz2a and Baz2b. *J. Med. Chem.* **2015**, *58*, 2553-2559.
125. Shen, D.; Wu, G.; Suk HI. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng.* **2017**, *19*, 221-248.
126. Brody H. Medical imaging. *Nature.* **2013**, *502*, S81-S81.
127. Naylor CD. On the Prospects for a (Deep) Learning Health Care System. *JAMA.* **2018**, *320*, 1099-1100.
128. Krizhevsky, A., Sutskever, I. and Hinton, GE. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, **2017**, *60*, 84-90
129. Esteva A.; Robicquet A.; Ramsundar B.; et al. A guide to deep learning in healthcare. *Nat Med.* **2019**, *25*, 24-29

130. Campanella, G.; Hanna, MG; Geneslaw, L.; et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med.* 2019, 25, 1301-1309.
131. Esteva, A.; Kuprel, B.; Novoa, RA.; et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* **2017**, 542, 115-118.
132. Liu, Y.; Jain, A.; Eng, C.; et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med.* **2020**, 26, 900-908.
133. Cho, SI.; Sun, S.; Mun, JH.; et al. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network. *Br J Dermatol.* **2020**, 182, 1388-1394.
134. Han, SS.; Park, GH.; Lim, W.; et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS One.* **2018**, 13:e0191493.
135. Hekler, A.; Utikal, JS.; Enk, AH.; et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur J Cancer.* **2019**, 118, 91-96.
136. Elmore, JG.; Barnhill, RL.; Elder, DE.; et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ.* **2017**, 8, 358:j3798
137. Piepkorn, MW.; Longton, GM.; Reisch, LM.; et al. Assessment of Second-Opinion Strategies for Diagnoses of Cutaneous Melanocytic Lesions. *JAMA Netw Open.* **2019**, 2:e1912597
138. Cupitt, J. pyvips. <https://github.com/libvips/pyvips>. Accessed 5 Oct **2018**
139. Bradski, G. et al. The OpenCV Library. *Dr. Dobb's J. Software Tools* **2000**
140. Jupyter Widgets Community. Ipywidgets. <https://github.com/jupyter-widgets/ipywidgets>
Accessed 6 July **2018**
141. Arganda-Carreras, I.; Sorzano, COS.; Marabini, R.; Carazo, JM.; Ortiz-de Solorzano C.; Kybic, J. Consistent and Elastic Registration of Histological Sections using Vector-Spline

Regularization. *Springer*, volume 4241/2006, CVAMIA: Computer Vision Approaches to Medical Image Analysis, **2006**, 85-95

142. Paszke et al. PyTorch: an Imperative Style, High-Performance Learning Library. *Curran Associates, Inc.* **2019**, 8024-8035

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:

Garrett Gaskins

5B9A855A5B3548D...

Author Signature

12/17/2020

Date