VALIDATION, VISUALIZATION, AND ANALYSIS WITH SLAP (SEMANTIC LINK ASSOCIATION PREDICTION): DATA FROM THE ANTIDEPRESSANT DRUG CLASS

Anand Kulanthaivel[1]

With assistance from:
David Wild[2]

I-571 Chemical Informatics
Final Project & Paper
30 April 2013

[1] School of Library and Information Science, Indiana University Bloomington 47405
[2] School of Informatics and Computing, Department of Cheminformatics, Indiana University Bloomington 47405

Validation, visualization, and analysis with SLAP (Semantic Link Association Prediction): Data from the antidepressant drug class

Abstract

Antidepressants (ADs) are a very diverse class of chemical compounds used in the treatment of Major Depressive Disorder (MDD), an affliction of the central nervous system (CNS). Furthermore, ADs represent a significant share of the pharmaceutical market, and are thus suited to analysis. SLAP (Semantic Link Association Prediction), a web application project by Chen et al, is designed to match chemicals to various gene/protein binding targets with a probability (p) value. SLAP predicts binding affinities semantically when no explicit binding study is found in its database. In this study, a luminance map visualization of SLAP scores between antidepressants and the proteins they are thought to target is created. It was found that Tanimoto clustering was a poor choice of array order in terms of visualization, and that a more coherent luminance map was created by performing a second-pass, confounded analysis on the original chemical order array with respect to SLAP predictions on their targets. This second-pass luminance map yielded more parsimonious conclusions, including the visualization of a "noradrenergic" pattern in drugs known to be noradrenergic. The second-pass clustering and visualization also more properly clustered selegiline and rasagiline, two MAO (monoamine oxidase) inhibitors, which were shown to have unusual activities via SLAP, not interacting with receptors but as expected, with the MAO enzyme.

Introduction: Use of SLAP

Due to the ambiguity of Ki values in the literature, the model we use is Semantic Link Association Prediction (SLAP)[1] in lieu of the Ki values. SLAP is a model that returns a p-value given any chemical entity and a gene-protein target. For example, sending SLAP the combination of duloxetine (Cymbalta; an MDD drug) with SLC6A4 (the serotonin reuptake transporter, one of the intended targets of duloxetine[2]) results in a SLAP p-value of 0.0004, which SLAP states is a strong binding association.

Choice of Evaluands

Chemicals to evaluate are chosen on a basis of diversity of putative action, but all chemicals are approved in the USA for therapy, either by itself or in combination with another medication, for Major Depressive Disorder as defined by the DSM-IV-TR.[3] Gene-protein targets were chosen from a broad array of receptors, transporters, and enzymes dogmatically known to be targeted by antidepressants and other CNS drugs. A list of targets being studied include:

(List of gene/protein targets studied; PubMed preferred names are used with common aliases listed in parantheses):

- SLC6A4 (SERT – Serotonin Reuptake Transporter)
- SLC6A2 (NERT/NET – Norepinephrine Reuptake Transporter)

[1] Chen, B. et al. (2012). "Semantic Link Association Prediction (SLAP)". http://cheminfov.informatics.indiana.edu:8080/slap/

[2] RxList.com (2004-2013). "CYMBALTA: Prescribing Information". [Monograph]

[3] American Psychological Association. (2000). "Diagnostic & Statistical Manual of Mental Disorders" 4th. Ed. [Text Revision].

- SLC6A3 (DAT/OATP – Dopamine Reuptake Transporter)
- HTR1A (5HT1A – Serotonin Receptor 1A)
- HTR2A (5HT2A receptor)
- HTR2C (5HT2C receptor)
- HTR3A (5HT3 receptor)
- DRD2 (D2 – Dopamine Receptor 2)
- ADRA2A (a2A – alpha-2A Adrenergic Receptor)
- ADRA2C (a2C – alpha-2C Adrenergic)
- MAOB (MAO – Monoamine Oxidase, Type B [CNS])

The chemicals to be studied are all approved in the USA by the Food & Drug Administration (FDA) for the treatment of Major Depressive Disorder. These chemicals are:

(List of chemicals; brand names are noted in CAPS in parentheses where known):

- Vilazodone (VIIBRYD)
- Escitalopram (LEXAPRO)
- Paroxetine (PAXIL; SEROXAT)
- Duloxetine (CYMBALTA; YENTREVE)
- Fluoxetine (PROZAC)
- Venlafaxine (EFFEXOR)
- Doxepin
- Bupriopion (WELLBUTRIN)
- Nortryptiline
- Selegiline (EMSAM)
- Rasagiline
- Mirtazapine (REMERON)
- Trazodone

Order of Array & Clustering Protocol

The first step was to determine an order of array for both the chemical/drug entities and the gene/protein targets. Clearly, an alphabetical order for either should not be entertained for this type of study. Therefore, drugs were initially clustered by pure chemical structure relationships as per PubChem Structure Clustering[4].

The classification of the gene-protein targets is a matter of even more serious debate: Classical mRNA or amino acid sequence homology can be used to order and cluster gene-proteins, but clustering in this fashion would be somewhat naïve, as there exists a wide homological discrepancy within the array of therapeutic targets of interest. In order to "humanize" the evaluation, then, the author has chosen to create a novel scheme involving subjective protein characteristics that clinicians may perceive as important in CNS disorders in general. Done as a multivariate binary clustering scheme by Euclidean distance in Minitab 16, the following "popular topics" are entertained regarding the medications:

- Does the protein foremost deal with X neurotransmitter in any way/shape/form? (Neurotransmitters being serotonin, dopamine, and norepinephrine for purposes of this project)

[4] National Institutes of Health (USA). (2013). PubMed: PubChem – Stucture Clustering. http://pubchem.ncbi.nlm.nih.gov/assay/?p=clustering

- What type of functional protein is it: e.g. G-protein coupled receptor, enzyme, transporter, or neurotransmitter-activated ion channel.

The exact clustering criteria of gene/proteins used can be seen in Appendix A. Note that as per the appendix, genes were classified in a pure binary scheme (e.g., SLC6A4/SERT scored '1' on serotonin activity, while SLC6A3/DAT scored '0', even though the latter will bind serotonin to some extent).

For the clustering analyses that would determine orders of array for both gene/protein targets and chemicals/drugs, Minitab 16 was used. In Minitab 16, all clustering was performed by Minitab's observational clustering function, single linkage, Euclidean distance measure.

Manipulation & Visualization

For automated visualization using existing software packages, the SLAP p-values were scaled and transformed into an order of magnitude figure (scaled SLAP P; SSP) by the following equation:

$$SSP = -ln(SLAP\ P)$$

Where "SLAP P" is the p-value returned by SLAP for any given association.

Such a transformation was required, as existing visualization frameworks such as Microsoft Excel and Sci2[5] (a network graph visualization tool) are, in terms of data input, not intended for such log-scale and order-of-magnitude differences; the human eye is entrained even less so.

Visualizations: First and Second Pass

In the second-pass study, the luminance map is replicated, but with the somewhat unsatisfactory visualization that was obtained (see figure in Appendix C), it was desired that the chemicals have a second-pass clustering so that there was some taxonomy given to them that took into account more than Tanimoto similarities. While the risk of confounding the variable is understood here, it is important to note that a parsimonious visualization is desired.

Second-Pass Quality Control: Pearson correlations

For a quantitative quality control assessment of the visualization, genes were compared *against each other* for similarities in SSP values across the order of second-pass clustered chemicals. This quality control analysis was visualized as a heat map and is seen in Appendix E. It is noted that higher correlation was seen close to and across the diagonal axis of the provided heat map, and that MAO formed an anti-correlation with just about every other target. However, it was observed that correlations were not in the expected pattern as one traverses the heat map towards the top-left side.

[5] Börner, K. et al. "Sci2" (Software).

Conclusions & Discussion

We conclude that for visualization and mapping purposes, that although pure Tanimoto similarity was a poor choice of evaluand for order of array for drugs/chemicals in the luminance map visualization. A confounding re-clustering using SLAP values obtained initially was re-performed in Minitab, and the new visualization allowed for greater conclusions to be drawn.

As we expect similar observations to cluster in darker or lighter clusters on a luminance map (or any other sort of graph), we observed a cluster of moderate activity through several norepinephrinergic drugs (including duloxetine, nortryptiline, mirtazapine, and trazodone); this phenomenon is seen in Appendix D. Furthermore, paroxetine was seen to have a generally heavy SLAP-scored interaction with many different receptors. Interestingly, by some clinicians' opinions, paroxetine is considered a side effect-heavy drug[6], and despite being officially classified as a purely selective serotonin reuptake inhibitor (i.e., an SLC6A4/SERT ligand and nothing else), it can be speculated that this particular compound has untoward effects on other targets. Since paroxetine is also the only antidepressant known to be teratogenic (toxic to the developing fetus in pregnant women)[7], its strong and broad effects could correlate with this clinical outcome.

Also, salient on the final luminance map were the similarities between selegiline and rasagiline, the two characteristic inhibitors of MAO (monoamine oxidase). It seems as if these two compounds, while being predicted by SLAP as being strongly affinitive the MAO enzyme, do not touch other targets. The contrapositive is also noted; i.e., other compounds seemed to stay away from MAO, providing a validation of the second-pass visualization.

Anomalies on the final luminance mapping that the author has yet to be able to explain include vilazodone's extreme putative affinity for the 5HT3A (HTR3A) receptor, which is not an affinity published by the manufacturer (instead, the manufacturer states that primarity affinity is for 5HT1A [HTR1A], which is demonstrated as much weaker than its affinity for 5HT3A).[8] Furthermore, an almost token 5HT3A ligand, mirtazapine, showed only borderline affinity for that target through SLAP. Whether this calls into question the manufacturer's studies or if SLAP is presenting an issue is a matter that would require further research.

Lastly, Pearson correlation to validate the orders of array present on the final luminance map also showed encouraging results. A heat map table of Pearson correlation by SLAP activity, again, is available in Appendix E. Since greater correlation is expected with greater gene/protein similarity, it is encouraging to see that darker values tended towards the center diagonal axis (the exceptions being anything related to MAO; this is also encouraging as MAO is obviously totally different from the other studied targets). In contrast, it is of interest that towards the upper-left of the Pearson heat map, there exist "bare spots" along the diagonal axis, and these spots correspond neatly to presence of the 5HT3A receptor as an evaluand. In the light of unexpected SLAP predictions with respect to this receptor and a couple of putative ligands as seen above, the idea of 5HT3A's misplacement on the map (or binding data derived for it) is in question.

---

[6] Baldwin D. & Birtwistle J. (1998). The side effect burden associated with drug treatment of panic disorder. *J. Clin. Psychiatry 59*(supp. 8): 39-44; 45-46.

[7] Einarson, A. (2012). Publishing statistically significant results with questionable clinical importance: Focus on antidepressant use in pregnancy. *J. Clin. Psychiatry. 73*(11):1443-1446.

[8] RxList.com (20011-2013). "VIIBRYD: Prescribing Information". [Monograph]

Future: Revision & Extension

The most immediate future for this project is adding bipartite network graph visualization, with drugs forming one party of nodes, chemicals forming another, and SSP values acting as the edge weights. Such a graph was attempted but found to be of limited utility due to a lack of visual comprehensibility. With the addition of human capita for visualization purposes (the information scientist is not exactly the best at art!), such a visualization in legible format may be possible.

Of course, anomalies (see conclusions section) must be further explained (and barring any explanation, the validity of one study or another may have to come into question.). Again, of particular interest for further study are anomalies seen on the 5HT3A (HTR3A) receptor with compounds vilazodone and mirtazapine. Further evidence for anomaly at the 5HT3A receptor is given by its relative juxtaposition in the Pearson correlation table (Appendix E). Thus, the 5HT3A receptor was the most anomaly-inducing target in terms of the visualization itself, so further research is certainly required on its order of array classification as well as binding studies performed on it.

Beyond that, it is clear that due to its small scale, the current project is simply a framework. There are only 11 protein targets and 14 drug chemical entities being studied; while the protein targets are probably almost the only ones studied classically in MDD, there are likely over one hundred drugs approved by the US Food & Drug Administration for treatment of MDD. The repertoire of chemicals could be extended yet further with a literature search for other CNS-related compounds that have shown efficacy in MDD but are not approved. Furthermore, the protein range can be extended to common CNS drug targets involved in pathogenesis of other CNS disorders (there are scores of such proteins).

Extension with further involvement in clinical and health informatics is also possible. Drug clustering may be re-done, for example, by drug side effects; also, gene-protein clustering could be done with measures of putative actions in different CNS disorders. Lastly, even the Pearson linear correlations obtained (see Appendix E) could be used to (albeit confoundingly) perform a third-pass re-ordering of array for another luminance map.

Appendix A: Gene-Protein clustering
Genes and their protein products/targets were clustered by the author's estimates of which neurotransmitters they dealt with and the nature of the protein's function.

The following table was obtained:

Name	GPCR	NA-IC	SLC	ENZ	5HT	NE	DA
SERT	0	0	1	0	1	0	0
NERT	0	0	1	0	0	1	0
DAT	0	0	1	0	0	0	1
5HT1A	1	0	0	0	1	0	0
5HT2A	1	0	0	0	1	0	0
5HT2C	1	0	0	0	1	0	0
5HT3	0	1	0	0	1	0	0
D2	1	0	0	0	0	0	1
a2a	1	0	0	0	0	1	0
a2c	1	0	0	0	0	1	0
MAO	0	0	0	1	1	1	1

Table. Binary values for classification of gene/protein targets.

(GPCR = G-protein coupled receptor; NA-IC = Neurotransmitter-activated channel; SLC = Solute carrier/transporter; ENZ = enzyme; 5HT = deals with serotonin; NE = deals with norepinephrine; DA = deals with dopamine).

The clustering obtained for genes and the key for the popular nomenclature used in the above table is on the next page.

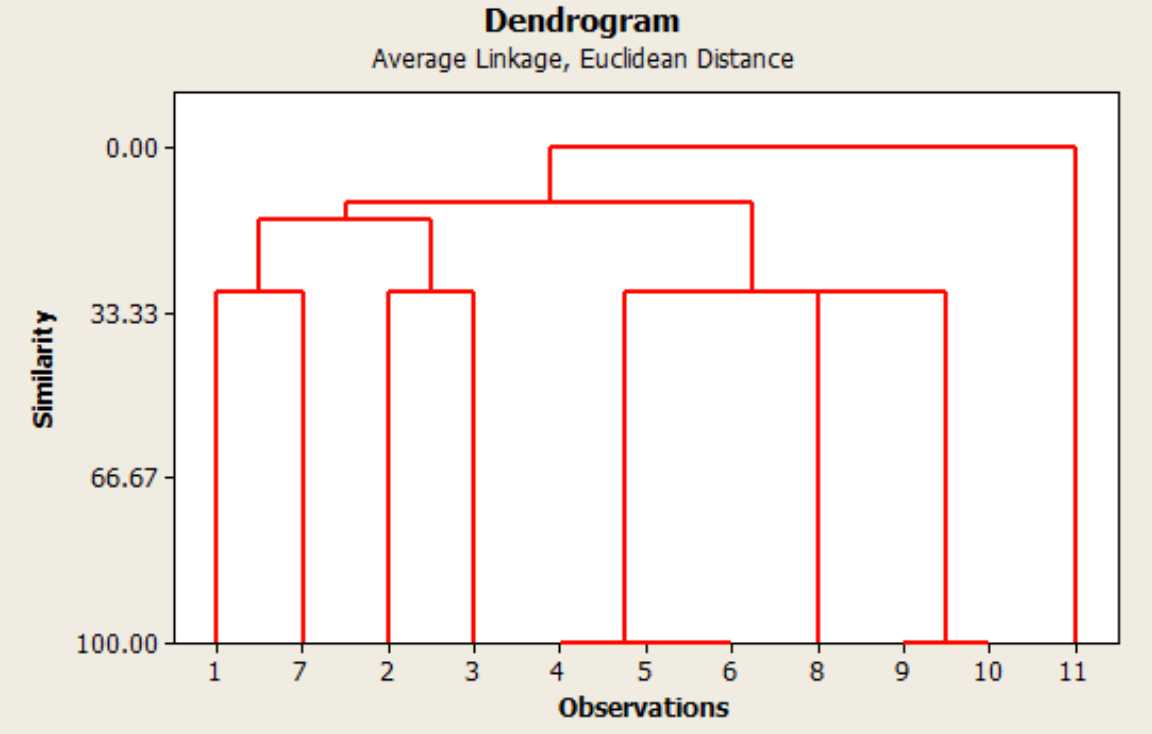

Figure. Cluster graph of gene/protein targets.

Appendix A Key (common names in parentheses):

1. SLC6A4 (SERT – Serotonin Reuptake Transporter)
2. SLC6A2 (NERT/NET – Norepinephrine Reuptake Transporter)
3. SLC6A3 (DAT/OATP – Dopamine Reuptake Transporter)
4. HTR1A (5HT1A – Serotonin Receptor 1A)
5. HTR2A (5HT2A receptor)
6. HTR2C (5HT2C receptor)
7. HTR3A (5HT3 receptor)
8. DRD2 (D2 – Dopamine Receptor 2)
9. ADRA2A (a2A – alpha-2A Adrenergic Receptor)
10. ADRA2C (a2C – alpha-2C Adrenergic)
11. MAOB (MAO – Monoamine Oxidase, Type B [CNS])

Appendix B: Chemical Clustering

One cluster set was obtained for pure Tanimoto clustering, but this graph will not be shown for the sake of brevity. The resulting order can be read on the bottom row of the luminance map provided in Appendix C.

However, after adjusting for SSP score (per target), keeping the same order of array of chemicals given by first-pass Tanimoto clustering, and performing reclustering, the following second-pass cluster graph was obtained:

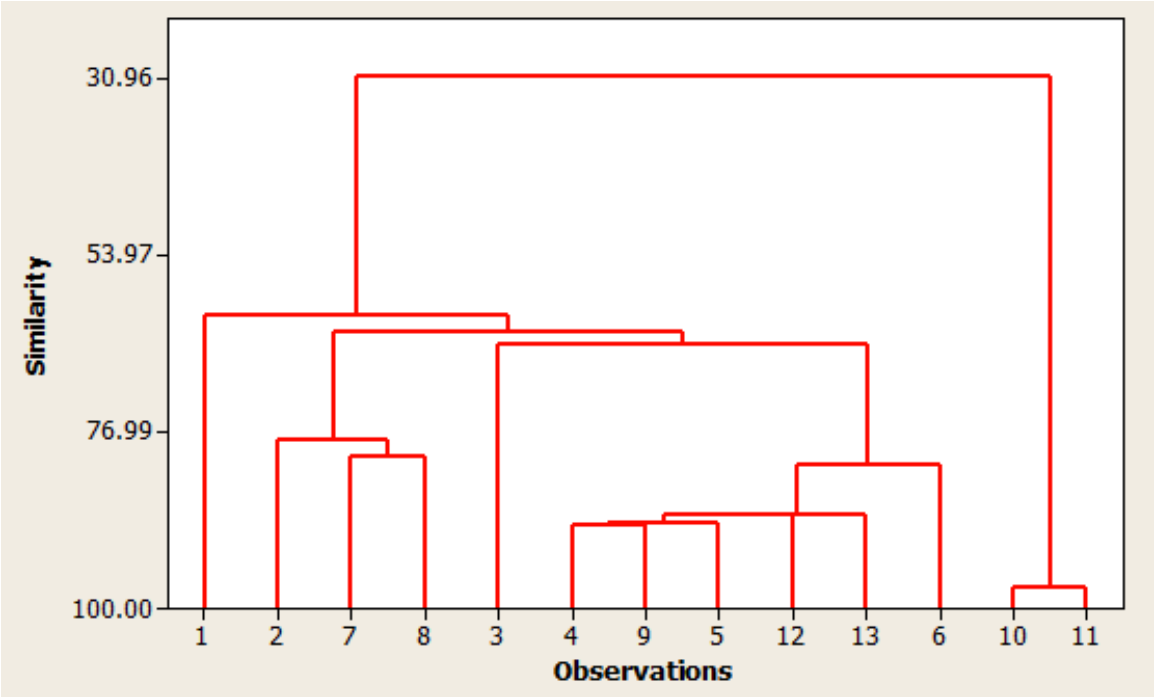

Figure. Second-pass clustering results for chemicals.

Key (BRAND NAME, IF KNOWN, IN CAPS):

1. Vilazodone (VIIBRYD)
2. Escitalopram (LEXAPRO)
3. Paroxetine (PAXIL; SEROXAT)
4. Duloxetine (CYMBALTA; YENTREVE)
5. Fluoxetine (PROZAC)
6. Venlafaxine (EFFEXOR)
7. Doxepin
8. Bupriopion (WELLBUTRIN)
9. Nortryptiline
10. Selegiline (EMSAM)
11. Rasagiline
12. Mirtazapine (REMERON)
13. Trazodone

Appendix C: First-pass luminance map of SSP values

Target	VZD	EPM	PXT	DLX	FLX	VFX	DXP	BPP	NTP	SLG	RSG	MZP	TZD
SLC6A4	7.1	10.9	10.5	7.8	9.2	8.1	11.5	10.9	8.5	3.4	3.4	9.8	8.1
HTR3A	17.6	5.3	6.8	7.8	6.4	7.8	5.6	5.5	8.5	2.1	1.7	6.9	6.5
SLC6A2	6.1	12.2	12.2	7.1	7.8	7.3	10.9	10.3	7.6	3.4	3.4	6.2	6.8
SLC6A3	6.3	10.3	12.3	7.1	7.8	8.1	13.2	10.4	7.0	2.8	2.9	6.2	6.8
HTR1A	10.2	7.6	10.9	9.6	10.0	9.6	7.4	6.3	11.1	1.9	1.6	10.4	9.2
HTR2A	10.5	10.8	11.6	10.4	10.6	10.3	7.3	6.9	11.0	4.2	4.1	10.4	9.9
HTR2C	9.9	5.6	11.4	10.4	10.6	10.3	6.6	5.4	11.2	4.2	4.1	9.9	9.8
DRD2	10.5	7.6	12.6	8.5	9.2	14.0	6.2	9.2	9.2	2.1	1.7	10.6	10.2
ADRA2A	11.3	5.7	6.6	10.4	10.3	10.3	6.6	5.0	10.6	4.6	4.5	11.0	8.5
ADRA2C	11.3	5.7	6.6	10.4	10.3	10.3	6.6	5.0	10.6	4.6	4.5	11.0	8.5
MAOB	4.4	5.4	3.6	3.6	4.7	4.0	2.5	2.9	4.1	10.1	10.2	4.5	3.3
Drug	*VZD*	*EPM*	*PXT*	*DLX*	*FLX*	*VFX*	*DXP*	*BPP*	*NTP*	*SLG*	*RSG*	*MZP*	*TZD*

Figure: First-pass luminance map. Note order of array of chemicals on the X-axis, which was done by pure Tanimoto clustering via PubChem). Compare to second-pass map in Appendix D.

Drug/Chemical Key:
VZD = Vilazodone; EPM = Escitalopram; PXT = Paroxetine; DLX = Duloxetine; FLX = Fluoxetine; VFX = Venlafaxine; DXP = Doxepin; BPP = Bupropion; NTP = Nortriptyline; SLG = Seligiline; RSG = Rasagiline; MZP = Mirtazapine; TZD = Trazodone.

For the final second-pass luminance map (next page), clustering hierarchy has been attached directly to the map itself in order for the reader to gain a better appreciation for the clustering families.

Appendix D: Final second-pass luminance map of SSP values

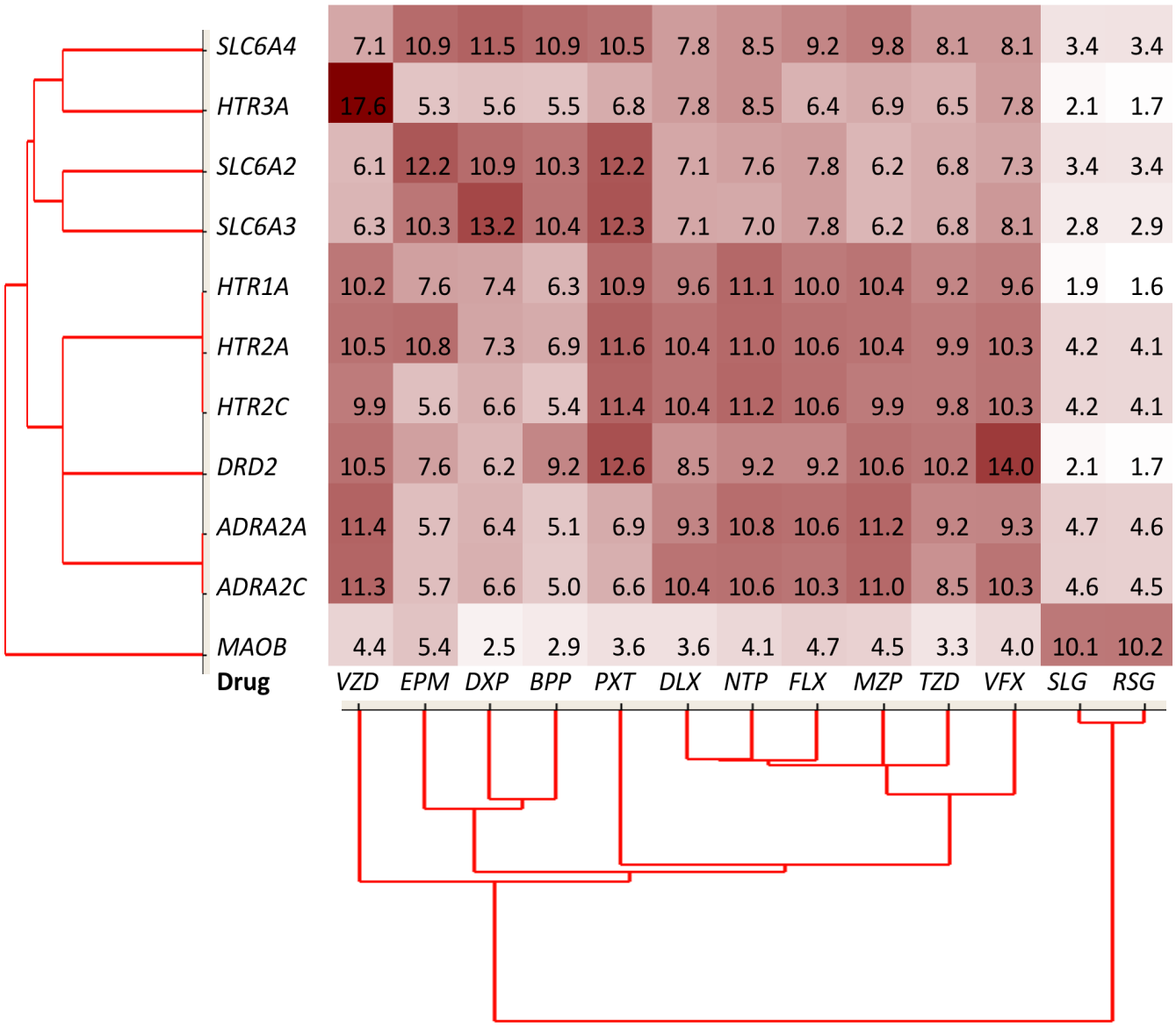

Figure. Second-pass luminance map with cluster families for both chemicals as well as gene/protein targets shown. SSP values are the same, but the order of array on the X-axis is enhanced. Note greater "color clustering" as compared to previous luminance map.

Drug/Chemical Key:
VZD = Vilazodone; EPM = Escitalopram; DXP = Doxepin; BPP = Bupropion; PXT = Paroxetine; DLX = Duloxetine; NTP = Nortriptyline; FLX = Fluoxetine; MZP = Mirtazapine; TZD = Trazodone; VFX = Venlafaxine; SLG = Selegiline; RSG = Rasagiline.

Appendix E. Pearson correlations between gene/protein targets.

Target	*SLC6A4*	*HTR3A*	*SLC6A2*	*SLC6A3*	*HTR1A*	*HTR2A*	*HTR2C*	*DRD2*	*ADRA2A*	*ADRA2C*	*MAOB*
SLC6A4	1	0.2003	0.8942	0.9025	0.6305	0.5966	0.3347	0.5901	0.1807	0.1658	-0.8457
HTR3A	0.2003	1	0.072	0.1167	0.6608	0.6065	0.5978	0.5864	0.7179	0.7117	-0.4915
SLC6A2	0.8942	0.072	1	0.9534	0.4455	0.4894	0.1688	0.4483	-0.1278	-0.1325	-0.671
SLC6A3	0.9025	0.1167	0.9534	1	0.4639	0.4314	0.2155	0.4788	-0.0745	-0.0665	-0.7592
HTR1A	0.6305	0.6608	0.4455	0.4639	1	0.9546	0.9223	0.8734	0.8093	0.7968	-0.8023
HTR2A	0.5966	0.6065	0.4894	0.4314	0.9546	1	0.8598	0.8385	0.7268	0.7198	-0.6934
HTR2C	0.3347	0.5978	0.1688	0.2155	0.9223	0.8598	1	0.8039	0.8517	0.8466	-0.615
DRD2	0.5901	0.5864	0.4483	0.4788	0.8734	0.8385	0.8039	1	0.6278	0.6394	-0.7671
ADRA2A	0.1807	0.7179	-0.1278	-0.0745	0.8093	0.7268	0.8517	0.6278	1	0.9819	-0.4715
ADRA2C	0.1658	0.7117	-0.1325	-0.0665	0.7968	0.7198	0.8466	0.6394	0.9819	1	-0.4758
MAOB	-0.8457	-0.4915	-0.671	-0.7592	-0.8023	-0.6934	-0.615	-0.7671	-0.4715	-0.4758	1

Figure. Heat map for Pearson correlation of genes in the author's order of array vs. their own SLAP activities. Note that this matrix is redundantly mirrored across the diagonal axis, with the central axis being 1:1 correlation, as the genes are identical in these cells.