Modeling Reactivity to Soft, Hard, and Biological Targets with a Deep Learning Network

Tyler B. Hughes,[*,†] Na Le Dang,[*,‡] Grover P. Miller,[*,‡] and S. Joshua Swamidass[*,†]

Department of Pathology and Immunology, Washington University School of Medicine, Campus Box 8118, 660 S. Euclid Ave., St. Louis, Missouri 63110, United States, and Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72205, United States

E-mail: tyler@wustl.edu; dangnl@wustl.edu; MillerGroverP@uams.edu; swamidass@wustl.edu

Unexpected drug toxicity is a critical problem for the pharmaceutical industry. Toxicity problems cause around 40% of drug candidates to be discontinued, oftentimes only after significant resources have been invested. Furthermore, drug-induced liver injury (DILI) is the most common reason already approved drugs are withdrawn from the marker, and causes half of all cases of acute liver failure, as well as 15% of all liver transplants within the United States.

Frequently, toxicity is caused by electrophilic drugs (and drug metabolites) that covalently bind to nucleophilic sites within biological macromolecules, including DNA and proteins. Conjugation to DNA can cause cancer, and conjugation to protein can cause a toxic immune response. For example, the well known hepatotoxicity of an acetaminophen overdose is due to metabolism of acetaminophen by Cytochromes P450 into the electrophilic metabolite *N*-acetyl-*p*-benzoquinone imine (NAPQI). NAPQI is electrophilically reactive and covalently binds to nucleophilic sites within proteins, resulting in hepatotoxicity in high doses (Figure 1).

[*]To whom correspondence should be addressed

[†]Department of Pathology and Immunology, Washington University School of Medicine, Campus Box 8118, 660 S. Euclid Ave., St. Louis, Missouri 63110, United States

[‡]Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72205, United States

Nucleophiles conjugate to specific atoms within reactive molecules: their sites of reactivity. Identification of a molecule's site of reactivity suggests the mechanism of its toxicity. Sites of reactivity encompass both harder electrophiles (such as saturated aldehydes) and softer electrophiles (such as alpha, beta-unsaturated aldehydes). Hard electrophiles generally react with hard nucleophiles, including purine and pyrimidine bases in DNA, whereas soft electrophiles tend to react with GSH and cysteine residues within protein. *In vitro* techniques to detect hard and soft electrophiles, such as incubation with the nucleophiles cyanide and GSH, are very expensive to perform for thousands of drug candidates. Furthermore, many of these assays do not yield site-level information, which can be sometimes be used to modify drugs to make them safer.

This work offers a computational alternative that can be used to quickly screen molecules for both hard and soft electrophilicity. From a literature-derived database, we extracted around 2000 reactions of drug-like molecules with cyanide, DNA, GSH, and protein. For each of these reactions, we labeled the exact site of reactivity on the starting molecule.

This data was inputted into a single deep neural network (Figure 2). From the 2D structure of each molecule, 208 topological descriptors are calculated in the input layer, which reflect various mathematical ways of describing each atom and molecule. Atom descriptors include traits such as atom identity or whether an atom is in a ring, and molecule descriptors include measures like molecular weight or hydrophilicity. All of these descriptors are inputted into the first hidden layer, which that produces reactivity scores for cyanide (ACRS in Figure 2), DNA (ADRS), GSH (AGRS), and protein (APRS). These atom reactivity scores range from 0 to 1, reflecting the probability of an atom conjugating to each nucleophile, and predict experimentally-observed sites of reactivity for cyanide, DNA, GSH, and protein with cross-validated accuracies of 96.9%, 90.2%, 93.4%, 94.4%, respectively. This approach outperformed individual models for each type of conjugation, due to the machine learning concept of transference learning.

The atom reactivity scores, as well as all molecule-level descriptors, are inputted into the second hidden layer, which calculates a molecule reactivity score (MRS in Figure 2). Across a structurally diverse dataset of aldehydes, benzoquinones, and esters, this MRS predicts the rate at which molecules react with glutathione with an R^2 of 0.91 (Figure 3). Site of reactivity data is an underutilized resource that can be used to not only predict if molecules are reactive, but also how they might be modified to reduce toxicity while retaining efficacy.

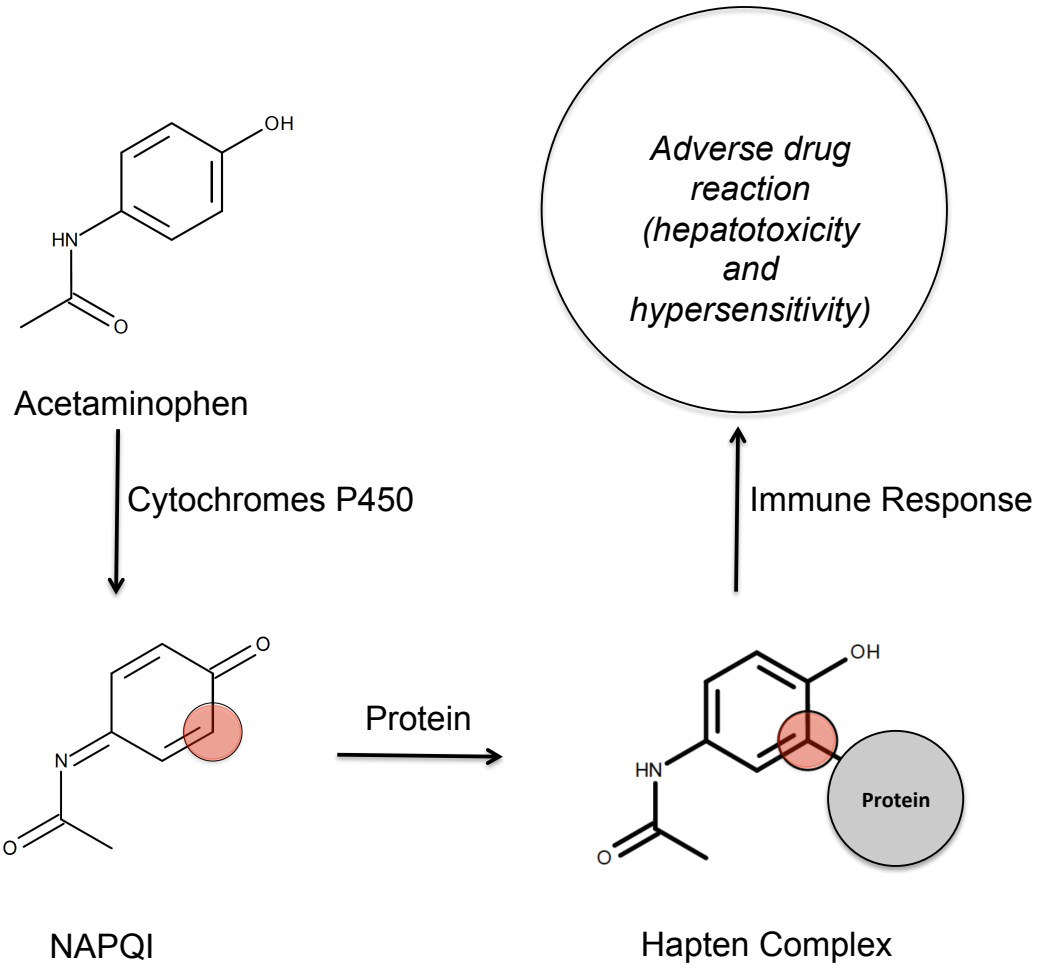

Figure 1: **Adverse drug reactions are often caused by electrophilic drug metabolites.**

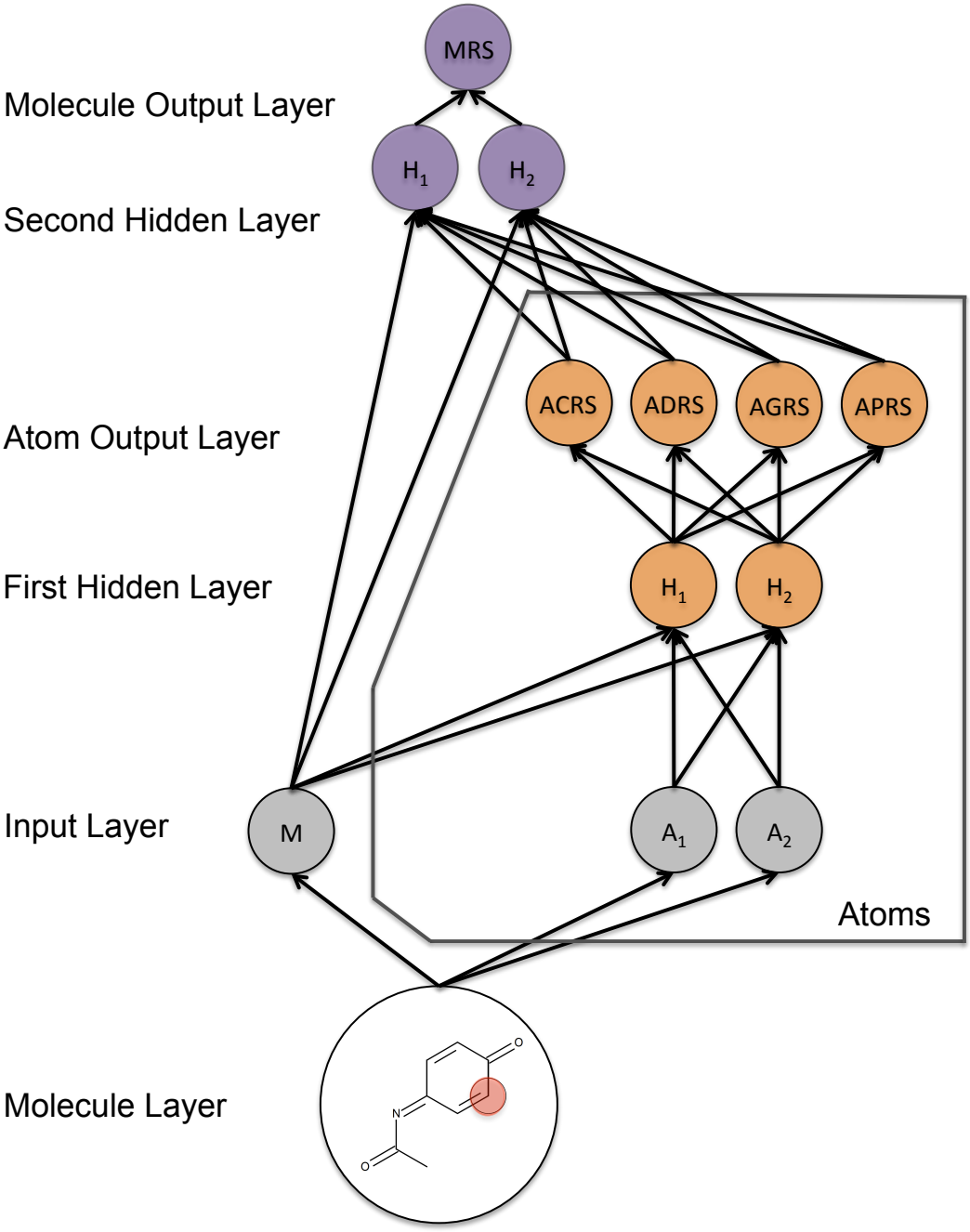

Figure 2: **The structure of the reactivity model.**

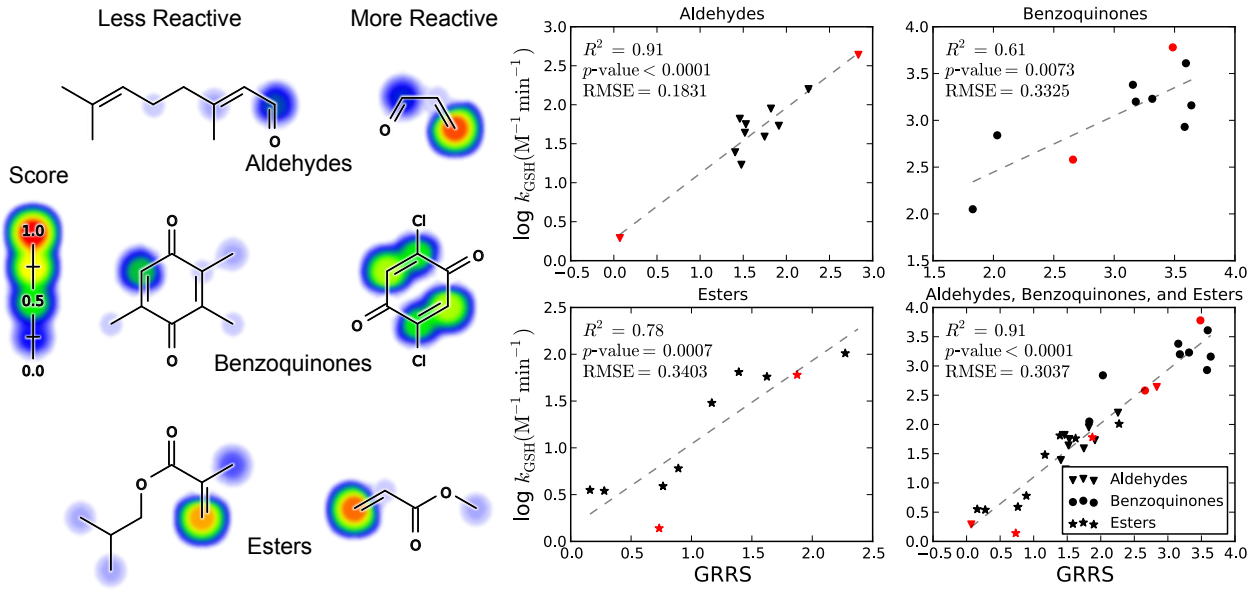

Figure 3: **XenoSite reactivity scores predict the GSH reactivity of aldehydes, benzoquinones, and esters.** In a series of three studies by Chan et al. (all published in J. Appl. Toxicol. in 2008), the GSH reactivity ($\log k_{\mathrm{GSH}}$) was determined for 11 aldehydes, 10 benzoquinones, and 10 esters. We used all 31 molecules as a training data set to model molecule reactivity. Each y-axis is $\log k_{\mathrm{GSH}}$, and each x-axis is the cross-validated molecule reactivity score (MRS). In addition to evaluating performance across the entire dataset, we also evaluated accuracy on each of the three datasets. Pearson correlation significance is indicated by the p-values. Six example molecules are visualized to the left, which correspond to the red data points on the right. The colored shading indicates atom-level GSH reactivity scores (which range from 0.00 to 0.92).