



LAWRENCE  
LIVERMORE  
NATIONAL  
LABORATORY

# Adverse Drug Reaction Prediction Using Scores Produced by Large-Scale Drug-Protein Target Docking on High-Performance Computing Machines

M. X. Labute, X. Zhang, J. Lenderman, B. Bennion, S. E. Wong, F. C. Lightstone

March 20, 2014

PLOS ONE

## **Disclaimer**

---

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

1 **Title:** Adverse Drug Reaction Prediction Using Scores Produced by Large-Scale  
2 Drug-Protein Target Docking on High-Performance Computing Machines

3  
4  
5

6 **Authors/Affiliations:** Montiago X. LaBute, Computational Engineering Division  
7 Xiaohua Zhang, Biosciences and Biotechnology Division  
8 Jason Lenderman, Computational Engineering Division  
9 Brian Bennion, Biosciences and Biotechnology Division  
10 Sergio E. Wong, Biosciences and Biotechnology Division  
11 Felice C. Lightstone, Biosciences and Biotechnology Division

12  
13  
14

15 **Institutional Affiliation:** Lawrence Livermore National Laboratory  
16 7000 East Avenue  
17 Livermore, CA 94550

18  
19  
20  
21  
22  
23  
24

25 **ABSTRACT**

26 Late-stage or post-market identification of adverse drug reactions (ADRs)  
27 is a significant public health issue and a source of major economic liability for  
28 drug development. Thus, reliable *in silico* screening of drug candidates for  
29 possible ADRs would be advantageous. In this work, we introduce a  
30 computational approach that predicts ADRs by combining the results of  
31 molecular docking and leverages known ADR information from DrugBank and  
32 SIDER. We employed a recently parallelized version of AutoDock Vina (VinaLC)  
33 to dock 906 small molecule drugs to a virtual panel of 409 DrugBank protein  
34 targets. L1-regularized logistic regression models were trained on the resulting  
35 docking scores of a subset of 560 compounds to predict 85 side effects, grouped  
36 into 10 ADR phenotype groups. Only 21% (87 out of 409) of the drug-protein  
37 binding features involve known targets of the drug subset, providing a significant  
38 probe of off-target effects. As a control, associations of this drug subset with the  
39 555 annotated targets of these compounds, as reported in DrugBank, were used  
40 as features to train a separate group of models. The Vina off-target models and  
41 the DrugBank on-target models yielded comparable median area-under-the-  
42 receiver-operating-characteristic-curves (AUCs) during 10-fold cross-validation  
43 (0.60-0.69 and 0.61-0.74, respectively). Evidence was found in the PubMed  
44 literature to support several putative ADR-protein associations identified by our  
45 analysis. Among them, several associations between neoplasm-related ADRs  
46 and known tumor suppressor and tumor invasiveness marker proteins were  
47 found. A dual role for interstitial collagenase in both neoplasms and aneurysm

48 formation was also identified. These associations all involve off-target proteins  
49 and could not have been found using available drug/on-target interaction data.  
50 The application of statistical analysis to highly parallelized molecular docking  
51 calculations and clinical databases presented in this study illustrates a path  
52 forward to comprehensive ADR virtual screening that can potentially scale with  
53 increasing number of CPUs to tens of thousands of protein targets and millions of  
54 potential drug candidates.

55

## 56 **INTRODUCTION**

57 Adverse drug reactions (ADRs) are detrimental, rare and complex  
58 perturbations of biological pathways by pharmacologically active small  
59 molecules. Each year ADRs cause 100,000 fatalities in the US[1]. One cost  
60 estimate of drug-related morbidity and mortality is \$177 billion annually[2], which  
61 is comparable to the public health burden of chronic illnesses like diabetes (\$245  
62 billion in 2012[3]). A systematic and accurate capability for reliably ruling out  
63 severe ADRs early in the drug development process currently does not exist. As  
64 a result, billions of research and development dollars are wasted as drugs  
65 present with serious ADRs either in late stage development or post-market  
66 approval. Highly publicized examples of phase IV failures include rosiglitazone  
67 (“Avandia”)[4] and rofecoxib (“Vioxx”)[5]. Early identification of serious ADRs  
68 would be ideal.

69 Although many ADRs are multi-factorial and depend on patient- and  
70 treatment-specific factors (e.g. genetic polymorphisms and medical history of the

71 patient, treatment dosages, environmental exposures, dynamics and kinetics of  
72 the relevant systems biology, etc.), all ADRs are initiated by the binding of a drug  
73 molecule to a target, whether these binding events are intended, on-target  
74 binding or promiscuous binding to one or more off-target proteins. Currently,  
75 pharmaceutical companies commonly employ experimental in vitro toxicity  
76 panels to assay small molecule binding to potentially critical protein receptors[6].  
77 Unfortunately, these panels probably do not include all of the proteins and  
78 receptors needed for high-accuracy prediction of serious ADRs[7]. Even if it were  
79 known how to augment toxicity panels to include a minimally complete set of  
80 receptors relevant for serious ADRs, there is uncertainty about how efficiently it  
81 could be screened.

82         An *in silico* platform that could accurately predict serious ADRs prior to  
83 costly in vitro screening panels and clinical safety trials is highly desirable and  
84 has been the focus of several recent studies.

85         A popular approach is to data-mine the publicly available databases for  
86 experimentally elucidated interrelationships between the chemical structures of  
87 drugs, their known interactions with proteins (most often their intended targets),  
88 and their known ADR profiles. An early study by Fliri and co-workers[8] clustered  
89 drugs based on their ability to inhibit a selected set of proteins. They showed that  
90 similar inhibition profiles indicated a similar set of side effects. More recently,  
91 Cobanoglu and co-workers[9] performed probabilistic matrix factorization on a  
92 1,413 drug x 1,050 known target protein matrix to learn a latent variable  
93 correlation structure between drugs and proteins. Drugs were then clustered in

94 this latent variable space, and it was found that drugs with similar therapeutic  
95 actions clustered together, independent of similarities in chemical structure. A  
96 highly cited effort by Campillos M. et al.[10] indicated that drugs with similar side  
97 effects have a correspondingly similar profile of protein targets. Another series of  
98 studies applied statistical machine learning approaches like support vector  
99 machines and sparse canonical correlation analysis (SCCA) to publicly available  
100 datasets to train models for ADR prediction. Pauwels et al.[16] used SCCA to  
101 relate PubChem[17] chemical substructure fingerprints of 888 approved drugs to  
102 1385 side effects in SIDER. Yamanishi and co-workers[18] used a similar  
103 approach to integrate drug-protein target data found in DrugBank and Matador  
104 with PubChem fingerprints to predict 969 SIDER side effects, applying both  
105 SCCA and a kernel regression method. They used the models to predict side  
106 effects in 730 previously uncharacterized small molecules in DrugBank where  
107 side-effect information was not available in SIDER. Finally, Liu et al.[19] found  
108 that adding phenotypic data on the drug (i.e. the presence or absence of side  
109 effects, excluding the one being predicted) to a similar feature representation to  
110 that considered in [18] greatly enhances prediction of the ADR of interest,  
111 obtaining AUCs > 0.9. However, since their approach relies on health outcomes  
112 data on the drug compound, the method is unsuitable for ADR prediction in the  
113 early-stage development of nascent drug compounds, prior to *in vitro* studies or  
114 clinical trials. In all of the cases listed above, only global quality-of-performance  
115 metrics, aggregated across all considered side effects, are reported, making it

116 difficult to assess how the models performed on individual side effects or classes  
117 of side effects.

118         There is another group of studies that more fully exploit the network  
119 structure of drug, protein, and ADR entity relationships. A network-oriented  
120 approach by Cami [20] analyzed a dataset consisting of 809 drug feature vectors  
121 (consisting of drug features from DrugBank and PubChem) and proprietary data  
122 on the drug side effect profiles. A unique aspect of the dataset is that the time  
123 ordering of when specific side effects appeared is reported. Starting with side  
124 effect profiles on the drugs from 2005, they trained a logistic regression model  
125 that could predict the side effects that manifested between 2006-2010,  
126 preserving the temporal order of how they manifest. The preservation of the time-  
127 ordering of the side effect appearance is appealing, but it is unclear how their  
128 approach would generalize to a different dataset. Mizutani[11] applied SCCA to  
129 find relationships between the drug-protein interaction network of 658 drugs from  
130 DrugBank and 1368 proteins extracted from DrugBank and Matador[12]  
131 databases to 1339 side effects associations as found in SIDER[13]. They found  
132 significant enrichment in most of the correlated protein-side effect sets for  
133 proteins involved in the same KEGG[14] and Gene Ontology biological  
134 pathways[15]. Similarly, Kuhn[21] constructed an explicit network to predict and  
135 characterize proteins that cause side effects by drawing statistical inferences  
136 between drug-target and drug-ADR links. Their method is able to reveal causal  
137 relationships between targets and ADRs but is highly sensitive to outliers. For

138 instance, there was insufficient statistical power to associate side effects to  
139 proteins that were an off-target of only a small number of drugs.

140         Indeed, the main weakness of these QSAR-like studies is their reliance on  
141 what is present in the experimental data, which will tend to feature a strong bias  
142 towards approved drugs (i.e. little representation of serious ADRs ) and on-target  
143 or intended effects. It is difficult to see how analysis of drug-intended target  
144 binding data could be applied to explore correlations between off-target drug-  
145 protein binding and possibly rare ADRs.

146         Recently, systems biology approaches have been used to predict ADRs  
147 by viewing ADRs as perturbations of biological pathways. These approaches  
148 seek to transcend the “one drug-one target” paradigm used in traditional drug  
149 design which ignores system-wide effects that cause a drug to have unforeseen  
150 pharmacological effects[22]. Scheiber et al.[23] integrated several chemical and  
151 biological databases by comparing perturbed and unperturbed pathways in a set  
152 of compounds that have a common toxicity phenotype. They use this analysis to  
153 link pathways with particular ADRs. Huang and co-workers[24] combined clinical  
154 observation data with drug-target data and the gene ontology (GO) annotations  
155 of the target proteins to predict ADRs. They find a significant improvement in the  
156 quality of their models by incorporating features from the protein-protein  
157 interaction (PPI) network of the targets. Similarly, Huang et al.[25] increased the  
158 median AUCs of their support vector machine models, from 0.591 to 0.700 by  
159 adding both PPI network and small molecule structural features to their feature  
160 set.

161 In all of these cited cases, the efforts to solve the ADR prediction problem  
162 have focused on integrating publicly available and (in some cases proprietary)  
163 biological (e.g. physical and chemical small molecule properties, drug-protein  
164 associations, protein-protein interaction networks, biological pathway and gene  
165 annotations, etc.) and epidemiological data on side effect-related health  
166 outcomes (e.g. FDA package label data, clinical trial data) to train statistical  
167 models to predict ADRs with various degrees for success.

168 A key drawback of using experimental data is that the type and quality of  
169 data that exists is influenced as much by the financial limitations of experimental  
170 drug development as by the relevant biological science. The drug-protein  
171 associations aggregated from DrugBank and Matador can be represented as a  
172 Boolean matrix where '1's ('0's) would indicate the presence (absence) of an  
173 association. This matrix has been used for some of the previous efforts, as noted  
174 above, and is highly sparse with '0's indicating both negative results of assays  
175 and unperformed assays. ADR-protein associations derived from these data limit  
176 us to patterns in known, intended "on-target" associations and limit the ability to  
177 find novel off-target associations. Also, data on lead compounds that have failed  
178 in the development pipeline are typically regarded as proprietary information and  
179 are generally unavailable for inclusion in analysis. Clearly, the majority of publicly  
180 available data is biased in ways that are difficult to correct for.

181 An alternative approach is to leverage ever-growing libraries of small  
182 molecule structures and databases of high-resolution experimentally solved  
183 protein structures, such as the Protein Data Bank (PDB)[26]. Technical advances

184 in drug-protein binding modeling, protein sequencing and homology modeling  
185 allow high-throughput virtual screening early in the drug discovery process. Vast  
186 libraries of small molecules can be docked to a large array of protein structures in  
187 order to simultaneously predict putative drug targets and ancillary, off-target  
188 binding interactions that may have associations to serious ADRs. Yang et al.[27]  
189 used virtual docking to propose possible interactions between a set of 845  
190 proteins and a set of 162 drugs that each must induce at least one of four ADRs.  
191 Lounkine et al.[28] predicted the activity of 656 marketed drugs on 73 targets  
192 from the Novartis in vitro safety panel using SEA. This was not a true docking  
193 study *per se*, in that SEA calculates the chemical similarity of each drug with  
194 each of the native ligands of the 73 targets. Similar to our current study, Wallach  
195 and co-workers[29] applied multiple stages of logistic regression to docking  
196 scores involving 730 drugs, 830 human protein targets and then applied multiple  
197 stages of logistic regression to this data and data on 506 ADRs, producing 32  
198 ADR-pathway associations supported by the scientific literature (i.e. PubMed).  
199 These studies used the “first principles” approach to circumvent the bias issues  
200 in experimental data outlined above, but none of these previous efforts describe  
201 computational frameworks scalable to the data sizes required for a high-  
202 accuracy, high-throughput ADR screening panel for nascent compounds.  
203 More recently, Reardon[30] reported on a computational effort that uses publicly  
204 available profiles of 600,000 chemical compounds and assesses their ability to  
205 bind to ~7000 chemical pockets on 570 human proteins. The known expression  
206 profiles of the proteins and receptors on human organs is then used to predict

207 where in the body a given drug will most likely take effect. While these efforts  
208 certainly operate at the necessary scale, they do not report a method to  
209 statistically associate the docking scores with ADR phenotypes, which is  
210 precisely the goal of our work here.

211 Our working hypothesis is that it is valuable to predict ADRs as early in the  
212 lead identification phase as possible. Structure-based, high throughput, virtual  
213 screening is already widely applied in the early stages of drug discovery because  
214 of its low cost and high efficiency in identifying putative drug targets. Molecular  
215 docking-based screening studies involve fitting a large library of  $N$  small  
216 molecules into the active sites of  $M$  target protein structures, to calculate  
217 estimates of binding affinities.  $M$  and  $N$  can be quite large. Currently, the PDB  
218 has  $M > 90K$  protein structures, increasing at a rate of over 7500 per year[26].  
219 The combinatorics of the possible chemical structural space occupied by small  
220 molecules is immense, recently estimated as  $N \approx 10^{60}$  possible drug  
221 compounds[31].

222 These numbers, combined with the complexities of conformational  
223 sampling to find the best fit of the small molecule (i.e. “pose”) in the target and  
224 the computational cost of the scoring function itself, make high-throughput ADR  
225 screening ideal for high-performance computing.

226 Zhang et al.[32] implemented a mixed parallel scheme using MPI and  
227 multithreading in the existing AutoDock Vina molecular docking program, called  
228 VinaLC. One million flexible docking calculations took about 1.4 hours to finish on  
229 ~15K CPUs. The docking accuracy of VinaLC has been validated against the

230 DUD (Directory of Useful Decoys) database by the re-docking of X-ray ligands  
231 and an enrichment study. The statistical results shown in Table 1 of [32] show  
232 VinaLC has a mean receiver operator characteristic area-under the curve (ROC  
233 AUC) of 0.64 (95<sup>th</sup> CI: 0.60-0.68) for the DUD set of decoys/ligands. Root mean  
234 square deviation (RMSD) values for self-docked ligands validation can also be  
235 found in [32]. As shown in Figure 4 of [32], 64.4% of the top scoring poses were  
236 identified with RMSD under the 2.0 Angstrom cutoff while that for the best poses  
237 is 70.0%. For the best poses, all the targets have RMSD values within 10  
238 Angstroms and about half of the targets have RMSD values less than 1  
239 Angstrom. Overall, the VinaLC docking program performed well for re-docking  
240 the X-ray ligands back into the active site of the X-ray structures with the default  
241 setting for the grid sizes and exhaustiveness = 8. A massively parallel virtual  
242 screening pipeline for Molecular Mechanics/Generalized Born Surface Area  
243 (MM/GBSA) rescoring has been developed to improve enrichment[33]. The  
244 MM/GBSA rescoring method improves the docking benchmark AUC to 0.71, on  
245 average. Overall the results demonstrate that MM/GBSA rescoring has higher  
246 AUC values and consistently better early recovery of actives than Vina docking  
247 alone.

248         A significant fraction of these molecules (e.g. drugs approved by the  
249 regulatory agency like the U.S. Food and Drug Administration) are annotated  
250 with known associated ADRs in public databases such as SIDER. As in the prior  
251 work we have cited, machine learning methods can then identify statistical  
252 associations between these ADR outcomes and patterns in drug-protein binding

253 as revealed by our VinaLC docking scores, and the results can be used to build  
254 predictive models so the probabilities of certain ADRs can be predicted for a  
255 nascent or theoretical small molecule drug candidate that may not have  
256 undergone *in vitro* or clinical trial testing.

257 This study potentially provides a technological and methodological path  
258 forward to large-scale, high-throughput, *in silico*, ADR comprehensive screening.  
259 Our results indicate that molecular docking performed with sufficiently detailed  
260 docking models on high-performance computers may provide reliable, cost-  
261 effective, comprehensive high-throughput screening of a drug candidate for  
262 binding across many known on- and off-targets to predict clinically important  
263 ADRs.

264

## 265 **MATERIALS AND METHODS**

### 266 **Dataset Creation**

267 We extracted 4,020 Swiss-Prot protein knowledgebase UniProt ID  
268 numbers (<http://www.uniprot.org/>) for proteins that were identified as drug targets  
269 in DrugBank as of October 12, 2012 (<http://www.drugbank.com/>). Mappings to  
270 587 experimental structures in the Protein Data Bank (<http://www.rcsb.org/pdb/>)  
271 (PDB) were obtained using the pdbtosp.txt file (Nov 2, 2013) from  
272 <http://www.uniprot.org/docs/pdbtosp> which links PDB ID numbers to UniProt IDs.  
273 A set of quality control rules were then applied (Supplementary Figure 1) which  
274 further reduced the list of proteins down to a final set of 409 experimental PDB  
275 structures. If multiple structures were given for the same protein, they were

276 selected by criteria in the following priority order: (1) human species, (2) crystal  
277 structure, (3) resolution (in Angstroms). This set of PDBs included 33 structures  
278 belonging to 16 UniProt IDs that are a subset of a larger consensus *in vitro*  
279 toxicity panel. This panel consists of 44 targets that were presented as a  
280 minimum *in vitro* toxicology panel from a collaboration of four major  
281 pharmaceutical companies[6]. The structures of 906 FDA-approved small  
282 molecule compounds in SDF format were obtained from the “Orange Book” of  
283 approved products[34]. Drugs that have more than 20 rotatable bonds were not  
284 included because most of them are natural products. The 3-D structures of target  
285 proteins and the small molecule compounds were then prepared for molecular  
286 docking calculations as described below.

287 A set of 85 side effects were selected from the SIDER database  
288 (<http://sideeffects.embl.de/>; extracted on November 26, 2012) because they were  
289 associated with high morbidity, high case fatality ratio, and/or the need for  
290 extended hospitalization. Individual side effects were grouped into higher-level  
291 health outcome groupings to reduce noise and provide signals at the organ or  
292 system level. Individual side effects were identified as lowest level terms in the  
293 medical dictionary for regulatory activities (MedDRA)[35]. Following the work of  
294 Huang and co-workers[25], the side effects of interest were grouped into ten  
295 MedDRA-defined system organ classes: (1) Neoplasms, benign, malignant, and  
296 unspecified (“neoplasms”), (2) Blood and lymphatic system disorders  
297 (“bloodAndLymph”), (3) Immune system disorders (“immuneSystem”), (4)  
298 Endocrine disorders (“endocrineDisorders”), (5) Psychiatric disorders

299 (“psychDisorders”), (6) Cardiac disorders (“cardiacDisorders”), (7) Vascular  
300 disorders (“vascularDisorders”), (8) Gastrointestinal disorders  
301 (“gastroDisorders”), (9) Hepatobiliary disorders (“hepatoDisorders”), and (10)  
302 Renal and urinary disorders (“renalDisorders”). A subset of 560 of the 906  
303 compounds in our docking score set were found to have associations to at least  
304 one of the 85 side effects we consider. The complete list of side effects by organ  
305 class is presented in Supplementary Table I. We produce a 560 x 10 drug-ADR  
306 matrix where a ‘1’(‘0’) indicates the presence (absence) of one or more side  
307 effects in the group.

308         At the end of the dataset creation stage, we have a total of 906  
309 compounds (560 with ADR associations), 409 proteins, and 10 outcome groups,  
310 comprising 85 severe side effects.

311         In order to compare the ADR prediction capability of “off-target” effects,  
312 obtained by the molecular docking calculations, with that of experimentally  
313 derived “on-target” drug-protein associations, a 560 drug x 555 target protein  
314 association matrix was extracted from DrugBank. More precisely, in order for a  
315 specific protein to be in the list of 555 proteins, it must be identified as a  
316 ‘Target’ in the DrugBank database of one or more of the 560 drugs in our  
317 dataset. The matrix is boolean-valued where a ‘1’(‘0’) indicates the presence  
318 (absence) of the association in DrugBank.

319

320 **Drug-protein target molecular docking calculations using VinaLC**

321 The 409 target protein structures retrieved from the PDB were processed for  
322 molecular docking calculations. The raw PDB files were processed by our in-  
323 house Protein Function Prediction (PFP) pipeline[36]. The structures of the  
324 protein targets were cleaned and protonated. “Cleaning” was defined by the  
325 following: alternate location “a” records for atoms were kept, and any ligands (i.e.  
326 atoms designated as ‘HETATM’ after the TER record in the PDB file that are not  
327 part of common ions) were deleted. Molecular modeling software (Schrodinger  
328 Inc.) was used to protonate the protein structure. In those cases where a known  
329 catalytic site was identified, the centroid coordinates for the active sites/binding  
330 sites of the protein targets were determined by CatSid[37], otherwise, these sites  
331 were determined by Sitemap[38]. A similarity to a known catalytic site was  
332 identified in 83 cases. Cofactors, metals, and crystallographic waters were  
333 removed from the protein structure when performing the docking calculation.  
334 Missing residues in the active site were reconstructed. For NMR structures that  
335 had multiple models, the first model was used. Similarly, for structures with  
336 residues having multiple positions, the first one was used. These pre-treated  
337 protein target structures were further processed by the in-house program,  
338 preReceptor[33]. The program preReceptor provides interfaces to integrate  
339 several external programs for target protein preparation. The preReceptor  
340 program firstly determines the dimensions of docking grids by utilizing the  
341 dms[39] and sphgen programs[40]. The dms program calculates the molecular  
342 surface of the target protein, and the sphgen program fills the active site of the  
343 target protein with spheres. The dimensions of docking grids were determined by

344 finding the distribution of sphere along the X-, Y-, and Z-axes. The cutoffs were  
345 set when the distribution of spheres changes drastically. In order to reduce the  
346 computer time, the target protein was cut by a radius of 30 Å centered at the  
347 centroid of active site because the dimensions of the active site usually range  
348 from 20 to 40 Å. A cutoff with a radius of 30 Å is sufficient for preReceptor to  
349 determine the grid dimension. The dimensions of the docking grids and centroid  
350 of active site were stored for docking calculation in the next step. The AMBER  
351 force field f99SB[41] was employed in the calculation for the receptor grid. Non-  
352 standard amino acids distant from the binding site were converted to alanine.  
353 Otherwise, non-standard amino acids were stored in the library, if present in the  
354 active site. Parameters for non-standard amino acids were calculated by AMBER  
355 antechamber. The energy minimization of the protein target was carried out using  
356 MM/GBSA[33] implemented in the sander program of the AMBER package[41].  
357 The structures were minimized with heavy atom constraints so the geometry of  
358 the active site remains unchanged. The PDB files of energy-minimized protein  
359 structures were converted to PDBQT files, which are used in the docking  
360 procedure. During the conversion, the non-polar hydrogen atoms are removed  
361 from the protein target structures.

362 The set of 906 approved drugs were processed by the in-house program,  
363 preLigand[33]. Similar to the program preReceptor, the program preLigand  
364 provides interfaces to integrate several external programs for ligand preparation.  
365 All drug compounds were parameterized using the AMBER GAFF force field as  
366 determined by the antechamber program in the Amber package[41]. Partial

367 charges of ligands were calculated using the AM1-BCC method. The structures  
368 of ligands were energetically minimized by the MM/GBSA[33] method  
369 implemented in sander. The atomic radii developed by Onufriev and coworkers  
370 (AMBER input parameter igb=5) were chosen for all GB calculations [42]. Those  
371 atoms with GB radii missing from the original program (i.e. fluorine, using a GB  
372 radius of 1.47 Å) were added into the sander program. The PDB files of energy-  
373 minimized ligand structures were converted to multiple-structure PDBQT files,  
374 which were used in the docking procedure. As with the receptors, non-polar  
375 hydrogen atoms were removed from the ligand structures. All these steps  
376 mentioned above have been integrated into the preLigand program.

377 The VinaLC parallel docking program[32] was employed to dock the 906 drug  
378 compounds into the 409 protein targets. [In our previous work \[\(JCIM DOI:  
379 10.1021/ci4005145\)\]](#), it was found that keeping 5–10 poses strikes a good  
380 compromise between accuracy and computational expense. For each of the 906  
381 x 409 = 370,554 individual drug-protein complex docking calculations, 20 poses  
382 were kept, Docking calculations used the coordinates of centroids and  
383 dimensions of active sites determined from the previous steps. The PDBQT files  
384 for target proteins and compounds obtained from previous steps were used as  
385 input files. The docking grid granularity was set to 0.333 Å. The exhaustiveness  
386 was set to 12, so that 12 Monte Carlo simulations search for docking poses for  
387 each complex. The whole calculation was finished within 1 hour on a high  
388 performance computer at LLNL using ~15K CPU cores. The top 20 docking  
389 poses were saved for each complex. The top docking score of each complex

390 were extracted from the docking results. A table of docking scores for the 906  
391 ligands X 409 receptors, together with compound's PubMed ID/name and protein  
392 PDB ID, was saved in the CSV format for the statistical analysis described in the  
393 following section. Finally, we constructed a virtual version of the consensus  
394 toxicity-screening panel of 33 protein receptors. For this smaller 560 x 33 subset  
395 of scores, MM/GBSA[43-50] rescoring calculations were performed on the Vina  
396 docking poses. To achieve high throughput, molecular docking programs usually  
397 employ the scoring functions that often use less computationally intensive  
398 methods, such as molecular mechanics force-field methods, empirical scoring  
399 functions, and/or knowledge-based potentials[50]. The scoring functions often  
400 simplify the calculation by neglecting important terms that are known to influence  
401 the binding affinity, such as, solvation, entropy, receptor flexibility, etc[51, 52]. A  
402 very popular practice is to rescore top-ranking docking poses using the more  
403 accurate, albeit computationally costly, MM/GBSA method to overcome  
404 shortcomings in the docking scoring function[33]. The MM/GBSA method  
405 accounts for the solvent and entropy effects more accurately. Solvation effects,  
406 mainly contributed by water molecules in the biological systems, play a critical  
407 role in ligand binding by providing bulk solvent stabilization and solute-  
408 desolvation, increasing the entropic contribution with the release of water  
409 molecules in the active site upon binding, serving as molecular bridges between  
410 the ligand and receptor[51].

411

## 412 **Statistical analysis**

413           The molecular docking calculations produced a 906 x 409 drug-protein  
414 docking score matrix. A 560 x 409 subset was extracted, where each of the 560  
415 compounds has at least one side effect, as reported in SIDER, for the 10 ADR  
416 groups we are considering. Statistical analyses was performed on this data to  
417 train predictive models of serious ADRs and characterize putative ADR-protein  
418 associations and is outlined below.

419           For the analysis, four separate data matrices are considered: (A) a 560 x  
420 409 VinaLC drug-protein docking scores (“Vina off-targets”) and (B) a 560 x 555  
421 DrugBank drug-target protein association matrix. Matrix (A) is used to train  
422 logistic regression models that allow off-target ADR-protein correlations to be  
423 explored. Matrix (B) is used to train models on “on-target” drug-protein  
424 associations. The comparison of results between matrices (A) and (B) enable  
425 comparisons to be made between the relative predictive capabilities of intended  
426 target proteins and off-targets across the different ADR groups. The 16 toxicity  
427 panel target proteins in isolation are considered, so we also have a (C) 560 x 16  
428 docking score matrix which is a subset of (A) and finally (D) a 560 x 16 boolean  
429 matrix which is analogous to (B), representing any drug-target associations  
430 reported in DrugBank between the 560 compounds and the 16 proteins of the  
431 toxicity panel. It is noted that the separate matrices (C) and (D) are constructed  
432 for the same on-target/off-target comparison purpose as matrices (A) and (B).  
433 Regarding the construction of the (C) matrix, there were 33 structures for the 16  
434 proteins, thus multiple PDB structures mapped to the same UniProt ID were  
435 averaged over, so (C) and (D) matrices are conformable. We note here that this

436 was only done for the virtual toxicity panel. For the main VinaLC docking score  
437 matrix (A), the scores for individual PDB structures were mapped one-to-one to  
438 the relevant UniProt ID for that protein. The elements in matrices (B) and (D) also  
439 correspond to single UniProt IDs.

440         Next we define thresholds so the docking scores in matrices (A) and (C)  
441 can be used as a heuristic for drug-protein binding. Global and protein-specific  
442 thresholds are defined. The raw docking score itself is used as a continuous  
443 feature, and (given that more negative scores correspond to stronger binding)  
444 additional thresholds are defined such that a docking score below the threshold  
445 indicates binding or, if above it, not binding. The docking score does not  
446 correspond to an actual energy, and it is difficult to set a single value for a  
447 threshold. Several thresholds are tried, letting the quality of the models (as  
448 quantified by the AUC) determine the best threshold for each ADR. For the Vina  
449 scores, ten feature sets are used, based on different choices of threshold: (1) raw  
450 docking scores, and then a series of global binding cutoffs: (2) -4.0, (3) -6.0, (4) -  
451 8.0, (5) -10.0, and (6) -12.0. Four additional thresholds based on protein-specific  
452 score percentiles were also defined: (7) 5th percentile, (8) 10th percentile, where  
453 the percentiles refer to the docking scores across all 560 compounds for a given  
454 protein. The last two thresholds were calculated by transforming the 560 docking  
455 scores for each protein into z-scores (i.e. transformed to have zero mean and  
456 unit standard deviation). Thresholds of (9) 1 standard deviation (SD) below the  
457 mean score (as used in the docking studies of Wallach and co-workers[29]) and  
458 (10) 2 SDs below the mean are also used. For the 560 x 16 virtual toxicology

459 panel, which used GBSA scores, the global thresholds were -15, -20, -25, -30,  
460 and these can be interpreted as binding free energies. Raw scores, protein-  
461 specific percentiles, and z-score thresholds are used as features, analogous to  
462 the thresholds defined for the VinaLC score matrix (A).

463 Logistic regression models were trained and selected through 10-fold cross-  
464 validation (CV) applied to the ten feature sets each for the data matrices (A) and  
465 (C) and then for the Boolean matrices (B) and (D). The training samples were  
466 labeled by the 560 x 10 response matrix, consisting of the Boolean associations  
467 between the 560 compounds and the ten ADR groups, leading to 22 separate CV  
468 runs in all.

469 The lasso penalty or L1 model regularization[53] is an effective method for  
470 continuous variable selection in the regime, where the number of training  
471 samples is comparable to (or may actually exceed) the number of training  
472 samples (i.e.  $p \approx n$  where  $p$  is the number of potential predictor variables, and  $n$  is  
473 the number of training samples). The L1 penalty term is proportional to the sum  
474 of regression coefficients  $|\beta|$  that fall off faster than the  $\beta^2$  terms used in L2  
475 regularization for small values of beta, so the lasso penalty is efficient at  
476 shrinking the betas to exactly zero, enabling sparse solutions and thus greater  
477 interpretability. The sparseness makes this method especially effective in the  
478 biological domain, where frequently a much smaller subset of the features are  
479 explanatory of the phenotype or outcome. L1 logistic regression has been

480 successfully applied to single nucleotide polymorphism (SNP) analysis[54], as  
 481 well as in previous ADR prediction studies[29].

482 The ADR prediction problem considered here can be formalized as a  
 483 case-control problem where a dichotomous variable  $y_{ki} \in \{0,1\}$  is defined for the  $i$ -  
 484 th sample and  $k$ -th ADR health outcome group with '1' coding cases and '0'  
 485 indicating controls. Given a feature vector for the  $i$ -th sample,  $\vec{x}_i$ , the probability  
 486 for the  $k$ -th outcome is given by

487

$$488 \quad p(y_{ki} = 1 | \vec{x}_i) = \frac{1}{1 + \exp[-\vec{\beta}_k^T \cdot \vec{x}_i]} \quad , \quad [1]$$

489 where  $\vec{\beta}_k = (\beta_{k0}, \beta_{k1}, \dots, \beta_{kp})$  is the parameter vector (including an intercept term) for  
 490 the  $k$ -th outcome, and is typically estimated by maximizing the log-likelihood  
 491 function

$$492 \quad L(\vec{\beta}_k) = \sum_{i=1}^n [y_{ki} \ln p_{ki} + (1 - y_{ki}) \ln(1 - p_{ki})] - \lambda \sum_{j=1}^p |\beta_{kj}| \quad [2]$$

493

494 where the second term in Eqn.(2) is the lasso penalty.

495 The L1-regularized logistic regression was used as implemented in the  
 496 glmnet package of Friedman and co-workers[55] in the 'R' statistical  
 497 programming environment. For each of the 10 ADR outcome groups in turn, one-  
 498 vs-all logistic regression was used with 10-fold cross validation. During 10-fold  
 499 cross validation, the following was done simultaneously: the objective function

500 (area under the receiver-operator characteristic curve (AUC)) was maximized,  
501 the model parameters in Eqns. (1) and (2) were estimated, and the optimal L1  
502 penalty parameter in Eqn.(2) was chosen as the one corresponding to the  
503 maximum median AUC. Each 10-fold CV was repeated ten times to average over  
504 sampling variability.

505 For each of the Vina off-target matrix (A) and the MM/GBSA off-target  
506 560x16 matrix (C), the feature set that had the best median AUC is selected and  
507 the “best model” is considered. For the DrugBank-derived data matrices (B) and  
508 (D), the best median AUC score was chosen as the best model.

509 The statistical significance of putative associations between the ADR  
510 groups and docking score matrix protein features were calculated. Statistical  
511 significance of the association for a putative ADR-protein pair was determined by  
512 the following procedure: univariate p-values for each ADR-protein pair were  
513 calculated using Fisher’s exact test if the protein feature was dichotomous (i.e.  
514 associated with a binding threshold, or DrugBank association). If the feature was  
515 continuous (i.e. the raw docking scores), the Wilcoxon rank sum test was used.  
516 In addition to p-values, we analyzed the false discovery rate (FDR) due to  
517 multiple hypothesis testing. For the models associated with the larger Vina off-  
518 targets matrix (A), we calculated q-values, using the ‘qvalue’ R-package of  
519 Storey[56], which gives us a way to manage the high false discovery rate that  
520 can be associated with large feature sets. For the smaller, virtual, toxicity,  
521 MM/GBSA matrix (C), the FDR was managed by applying a simple Bonferroni  
522 correction[57] to the p-value.

523 The workflow just described, comprising data integration between  
524 DrugBank, UniProt, the PDB, and SIDER, as well as our docking score  
525 calculations and subsequent statistical analyses, is shown schematically in  
526 Figure 1.

527

### 528 **PubMed Text Mining to find supportive evidence of ADR-protein** 529 **associations**

530 PubMed database (<http://www.ncbi.nlm.nih.gov/pubmed>) queries were  
531 used to search for evidence in the literature to support putative ADR-protein  
532 relationships identified by the statistical analyses of the VinaLC drug-protein  
533 docking matrix. The protocol for searching the PubMed database was as follows:  
534 1) Queries for co-occurrences of the UniProt name of the protein and the  
535 MedDRA lowest-level term (LLT) of each individual side effect constituent of the  
536 ADR group were performed, 2) If the number of hits returned was substantive  
537 (~10), or the quality of the hits was high, then the association was triaged for  
538 manual review of the PubMed results set. An example of a high quality hit is the  
539 side effect and the protein terms co-occurring in the title or abstract of an article.  
540 ADR-protein associations that passed the manual review process were deemed  
541 significant and included in Tables I and II.

542

## 543 **RESULTS**

544

545           The 560 x 10 drug vs ADR group matrix (C) and the 560 x 409 drug vs  
546 protein docking score matrix (A) were used to train logistic regression models  
547 using L1-regularization, which allows the model to focus on high-information  
548 predictors and helps reduce over fitting. Figure 2 presents the performance  
549 profile of our ADR prediction models. For each ADR group, a “best model” was  
550 chosen based on the median AUC score of a model obtained during a single ten-  
551 fold cross-validation run. The quality of these models was compared to models  
552 trained on the 560 drug x 555 DrugBank protein target matrix (B), using the  
553 identical statistical model training procedure that was applied to the 560 x 409  
554 VinaLC docking score matrix (A). Figure 2 also compares the performance  
555 profile of the docking score models with that of the models trained on the  
556 DrugBank data. Across all ADR groups, the range of the best model AUCs for  
557 the VinaLC “off-target” models was 0.60-0.69. The corresponding AUC range for  
558 the DrugBank “on-target” models was AUC=0.61-0.74. Focusing on single ADRs,  
559 the inter-quartile range of the VinaLC “off-target” AUCs are above those of the  
560 DrugBank “on-target” models for both ‘neoplasms’ and ‘vascularDisorders’ ADR  
561 groups. The AUC distributions are not significantly different between the two  
562 datasets for ‘immuneSystem’ and ‘bloodAndLymph’. The DrugBank model AUCs  
563 were larger for these ADR groups: ‘psychDisorders’, ‘endocrineDisorders’,  
564 ‘renalDisorders’, ‘hepatoDisorders’, ‘gastroDisorders’, and ‘cardiacDisorders’.  
565 The difference in AUCs implies the importance of the on-target binding  
566 contributions for this subset of ADRs.

567           The ability of docking score data to identify potential associations between  
568 off-target drug-protein binding and individual side effects in the ADR groups were  
569 investigated. Additional statistical analysis was performed on the VinaLC drug-  
570 protein docking score matrix and the logistic regression models to derive  
571 associations between ADR groups and proteins. Only 21% (87 out of 409) of the  
572 drug-protein binding features involve known protein targets of the drug subset,  
573 providing a significant probe of off-target effects. In Table I, side-effect protein  
574 pair-wise associations are shown rank-ordered in ascending order, according to  
575 that feature's p-value. For each entry we list the UniProt name and ID of the  
576 drug-binding protein, the PDB ID for the protein target used in docking, the p-  
577 value, the corresponding q-value to indicate the FDR for that feature, and the  
578 beta coefficient in the "best" model. Furthermore, the variable selection capacity  
579 of L1-regularization was employed, so that a protein feature must have a non-  
580 zero beta coefficient in order to have been included in Table I. Finally, for  
581 inclusion in Table I, the ADR-protein association needed to pass the manual  
582 review of PubMed evidence. In the last column of Table I, the level of evidence  
583 from PubMed that supports the ADR-protein correlations is shown. For a specific  
584 putative ADR-protein entry in Table I, counts in parentheses show the number of  
585 papers found in PubMed that contain the co-occurrence of (1) the MedDRA  
586 lowest level term for a component individual side effect from the ADR group and  
587 (2) the UniProt name for the protein.

588           The associations between the ten ADR groups and a subset of the full  
589 VinaLC "off-target" docking score matrix (C) were investigated. Models trained

590 on a 560 x 16 subset of the full VinaLC docking score matrix (C) were compared  
591 to the models trained on a 560 x 16 DrugBank on-target subset matrix (D). The  
592 docking calculations were refined using the more computationally expensive and  
593 more chemically accurate MM/GBSA-correction of the Vina score. The same  
594 logistic model training procedure used on the larger predictor sets to train logistic  
595 regression models was applied to these smaller matrices. The boxplots of the  
596 “best model” AUCs for the screening panel models are shown in Fig.(3). Overall,  
597 the range of AUCs for the MM/GBSA “off-target” version of the consensus panel  
598 (AUC=0.55-0.65) and the DrugBank “on-target” version (AUC=0.58-0.69) of the  
599 panel indicate that the quality of the models are only marginally poorer than  
600 those derived from the larger predictor set, but use a factor of ~26 fewer protein  
601 features, indicating they may have some value in the drug development pipeline.  
602 Across the ADR groups, the MM/GBSA and DrugBank virtual panel model AUCs  
603 are similar for ‘immuneSystem’, ‘cardiacDisorders’, ‘gastroDisorders’,  
604 ‘bloodAndLymph’, and ‘hepatoDisorders’. The MM/GBSA-derived models for the  
605 ‘endocrineDisorders’, ‘psychDisorders’, and ‘renalDisorders’ ADR groups are all  
606 significantly worse than the corresponding DrugBank models.

607         Given the role of these 16 proteins in *in vitro* toxicity panels, it is of interest  
608 to see what specific associations they may have with specific side effects.  
609 Potential ADR-protein associations are shown in Table II, listed by UniProt name  
610 and ID. All potential ADR-protein associations had to have a Bonferroni-corrected  
611 p-value < 0.05 and a non-zero beta coefficient in the “best” logistic regression

612 model. Additionally, the associations had to pass the same manual review  
613 process used for the associations listed in Table I.

614

## 615 **DISCUSSION**

616         The major contribution of this work is a demonstration of the feasibility to  
617 holistically treat the ADR prediction problem for nascent drug compounds. Our  
618 methods treat the problem from microscopic levels (i.e. drug-protein binding) all  
619 the way up to prediction of clinical ADR phenotypes. We show, for our particular  
620 set of 560 drugs, that using molecular docking scores yields ADR prediction  
621 models comparable in quality (as evaluated by AUCs) to models developed using  
622 publicly available, experimentally-derived drug-protein associations. However,  
623 the AUCs, for both docking scores and experimental data, are not of sufficient  
624 quality for clinical prediction, and it is interesting to note the quality is poorer for  
625 highly multi-factorial disorders (e.g. cardiac disorders). As an example, for the  
626 virtual toxicity panel model quality results shown in Figure 3, we can see that for  
627 psychological disorders, the on-target relationships in the virtual panel yields a  
628 model with AUCs close to 0.7, while the MM/GBSA-rescored docking scores,  
629 emphasizing off-target effects, yields an AUC slightly better than random (i.e.  
630 AUC=0.5).

631         We first discuss some issues related to the molecular docking score  
632 calculations. The current work is focused on the binding of drug ligands to off-  
633 target proteins, where typically little or no data exists to inform initial placements  
634 of water molecules, metal ions, co-factors, and other hetero atoms. Additionally,

635 few if any parameters are available for these atom types in most docking codes.  
636 Given this and the lack of experimental or theoretical justification to guide  
637 placement of these atoms into the active sites our computational methods have  
638 predicted, we adopted the common practice for docking calculations, which is to  
639 remove the hetero atoms in the active sites [NEED CITATIONS]. This choice is  
640 consistent with state-of-the-art docking calculations, for example current virtual  
641 high-throughput screening leaves water molecules out as a rule [NEED  
642 CITATIONS]. Implicit solven was included in the MM/GBSA calculations.

643 As stated in the Methods section, we tried several different binding  
644 thresholds for the docking scores. Both the VinaLC and MMGBSA logistic  
645 regression model AUCs did not monotonically vary with choice of thresholds, and  
646 there were no clear trends with threshold choice with the exception that. The  
647 correlation between threshold choice and AUC was noisier in the ADR groups  
648 with lower overall AUC values, where the maximum value was, in some cases,  
649 only greater than the second largest AUC by a few  $\sim 0.01$ . For ADR groups with  
650 the best AUCs, the maximum AUC was often more clearly differentiated from the  
651 AUCs of the other competing threshold values.

652 Models trained to predict side effects in the ‘neoplasms’ and  
653 ‘vascularDisorders’ ADR groups on the full 560 x 409 VinaLC docking score  
654 matrix (A) perform better than their DrugBank-derived counterparts (B).  
655 We identify several potential off-target ADR-protein associations that would be  
656 impossible to find using only binding data between a drug and its intended  
657 protein targets (see Table I). Some of the more compelling associations found

658 are described below, along with supporting evidence from the literature. The  
659 literature cited here may describe examples where biological mechanisms are  
660 perturbed by drug binding to protein constituents of pathways associated with the  
661 ADRs.

662 **Interstitial collagenase (MMP1) with both neoplasms and vascular**  
663 **disorders.** Increased MMP-1 gene expression appears to be a biomarker for  
664 cancer metastasis. Specifically, we find evidence for separate constituents of the  
665 'neoplasms' group: breast neoplasms[58], adenocarcinoma[59], and glioma[60].  
666 Interstitial collagenase also seems to contribute to aneurysms. Specifically, cell  
667 distribution differences of MMP-9 and the tissue inhibitor of MMP-1 in patients  
668 with Kawasaki disease[61]. This work implicates interaction of MMP-9 and MMP-  
669 1 with aneurysm formation in Kawasaki disease.

670 **Tyrosine kinase Syk with breast neoplasms and adenocarcinomas.**  
671 A possible mechanism of interaction may be a role in suppression of breast  
672 cancer metastasis to lymph nodes[62], as well as regulating cell-cell adhesion  
673 and motility[63]. Some data suggest that Syk expression in the spleen may  
674 inversely correlate with the proliferation and invasive capacity of breast  
675 cancer[64]. Syk acts as a pancreatic tumor suppressor in pancreatic  
676 adenocarcinoma tumors, regulating cellular growth and invasion[65].

677 **Complement C3 with breast neoplasms.** An analysis[66] of expression  
678 patterns for acute phase proteins in breast, colorectal, and lung cancer indicate  
679 that the most accurate candidate biomarker for breast cancer in their panel was

680 Complement 3 (C3) as used in a univariate logistic regression model (AUC=0.89  
681 and 73% correct classification performance in leave one out cross-validation).

682 **Cytotoxic T-lymphocyte protein 4 (CTLA-4) with sarcoidosis.** A case  
683 study[67] shows exacerbation of sarcoidosis in a melanoma patient treated with  
684 anti-CTLA-4 monoclonal antibody inhibitor ipilimumab. Another study[68] reports  
685 correlations of specific CTLA-4 gene polymorphisms in sarcoidosis patients with  
686 different disease phenotypes.

687 **Profilin-1 with endocrine-related disorders.** Profilin-1 expression is  
688 markedly elevated in the atherosclerotic plaques of diabetics, showing a potential  
689 role in mediating diabetic-related vascular endothelial cell dysfunction[69].

690 **Coagulation factor IX with thyroid disorders.** A meta-analysis[70]  
691 looked at 29 trials and 11 studies and concludes that subclinical hyperthyroidism  
692 induces a pro-thrombotic state. More precisely, thyrotoxicosis shifts balance to a  
693 pro-coagulant/hypofibrinolytic state.

694 **Caspase-3 with bipolar disorder and schizophrenia.** Some papers  
695 hypothesize that enhanced cellular apoptosis is a disease mechanism in  
696 neurodegenerative diseases. A postmortem study on bipolar disorder patients  
697 shows significant increases in pro-apoptotic factors (inc. Bax, BAD, caspase-9  
698 and caspase-3)[71]. A population of anti-psychotic medicine-naive first-episode  
699 schizophrenia patients show higher caspase-3 activity and lower BCL2  
700 expression[72].

701 **Integrin beta-2 and myocardial infarction.** Studies have shown integrin  
702 and monocyte migration to ischemic myocardium. A study[73] that performed

703 flow cytometry-based whole-blood assays in 87 patients with unstable angina  
704 finds that beta-2 integrin mediated T-cell recruitment in coronary plaques  
705 identifies high-risk patients with severe coronary artery disease but no  
706 myocardial infarction and is predictive of future CV events even in the absence of  
707 myocardium damage markers like troponin or high-sensitivity C-reactive protein.

708 We also find ADR-protein associations for the 16-protein consensus  
709 panel. The protein targets are included in panels used by major pharmaceutical  
710 companies for *in vitro* screening of ADRs for drugs in the development pipeline.  
711 Our results provide a rationale, founded on independent calculations, for their  
712 inclusion in the panel based on side effect phenotypes for which they probe.  
713 Potential ADR-protein associations, supported by some level of evidence in  
714 PubMed, are listed in Table II. Among them, we found a correlation between  
715 agranulocytosis and the histamine H1 receptor (an example is the drug clozapine  
716 an H4-receptor agonist with some H1 activity)[74]. Also, a number of cardiac-  
717 related side effects were associated with Prostaglandin G/H synthase 2  
718 (Cyclooxygenase 2), in particular ‘myocardial infarction’ which yielded 217  
719 PubMed hits.

720 Using molecular docking scores for drug-protein matrices has advantages  
721 over other approaches to predict association of “off-target” effects. Molecular  
722 docking is a first-principles approach based on a physics-derived force field, such  
723 that only the structure of the drug and the protein are necessary. Not surprisingly,  
724 the docking approach does not have as strong a dependence on the availability  
725 of drug-protein correlations in manually curated biological or chemical databases

726 (which are biased toward intended, on-target effects), though this data can be  
727 integrated into our type of analyses as well. Experimental drug-protein  
728 association matrices are extremely sparse, i.e. there are large areas of the drug x  
729 protein matrix that are unexplored by *in vitro* assays or clinical trials. In contrast,  
730 the docking calculations enable an exhaustive probing of binding associations  
731 through the entire drug x protein matrix, allowing the exploration un-intended (i.e.  
732 off-target) interactions that might not have been previously experimentally  
733 investigated during drug development. Thus, docking scores provide a direct way  
734 to probe off-target effects.

735 Here we compare our work to previous efforts that have applied molecular  
736 docking to study ADR-protein correlations. A recent large-scale drug-protein  
737 docking exercise was described in [30], but this effort had a different goal than  
738 our study. While the work outlined in [30] appears to focus on a highly automated  
739 method where structures are prepared and docked in a bulk fashion, we have  
740 chosen to initially focus on a smaller group of drug-protein interactions, hand-  
741 curating the initial docking structures, so the quality of the drug-protein binding is  
742 sufficiently high that we can link to ADR outcomes downstream of docking. No  
743 attempt, beyond identifying the tissue tropism of the receptors used in docking, is  
744 made to correlate the results of docking to ADR phenotypes. The work of  
745 Wallach et al [29] bears some similarities to our work and here we list some of  
746 the major differences between the two efforts. Specifically, we: 1) use q-values to  
747 correct for multiple hypothesis testing, which has been previously shown to  
748 indicate “interesting” protein-side effect correlations[21], 2) focus on proteins

749 rather than pathways and on only on a small set of serious ADRs, 3) consider  
750 multiple binding thresholds for binding, in addition to 1-SD above the mean of the  
751 z-scored docking scores used by Wallach et al., 4) compare model performance  
752 across ADRs (thru AUC), where the work in [29] is focused on ADR-pathway  
753 associations, and 5) are interested in ADR prediction using the docking scores.  
754 Although, Wallach et al also use L1-regularization to mitigate over-fitting, our  
755 lambda parameter is chosen thru 10-fold cross-validation, while their lambda  
756 parameter seems to have been arbitrarily chosen to be  $\frac{1}{2}$  the value needed to  
757 suppress all beta coefficients to zero. They do not appear to discuss the  
758 associations they produce in quantitative terms (AUCs of the models, p-values of  
759 the associations). Also, their study treats each side effect individually, which may  
760 lead to bad class imbalances with more rare ADRs, a common problem in QSAR  
761 studies. We mitigate this issue by classifying ADR phenotypes into groups.

762         The limitations of our method can be categorized into two areas: 1)  
763 molecular docking and 2) ADR phenotypes. In the 'Introduction' section, we  
764 noted the inherent biases in the QSAR-like studies given their reliance on  
765 experimental data derived from approved drugs. While the molecular docking  
766 studies advocated here does not suffer the same bias towards approved drugs,  
767 the methods presented are biased heavily to proteins that have available of 3D  
768 structures, which restricts these methods to ~50% of the human proteome, as  
769 estimated by Xie et al. [CITE Xie paper]. Unfortunately, this missing cohort of  
770 proteins will be highly enriched with some of the more important classes for  
771 ADRs namely membrane-bound receptor proteins. Future work on our virtual

772 drug-protein panel will include a focus on applying state-of-the art methods to  
773 create and validate homology models to mitigate this bias.  
774 For molecular docking to be a feasible method for predicting “off-target”  
775 associations, the execution of the docking needs to be fast and reliable. Our  
776 implementation of the well-vetted Vina docking program, VinaLC, has been  
777 optimized for HPC and has been benchmarked with known limitations (e.g.  
778 metalloproteins)[32]. However, we are limited by the availability of 3D structures  
779 of target proteins relevant for side effects and the quality of those 3D structures.  
780 With the growing number of protein crystal structures and the higher quality  
781 homology models, we believe the availability of quality 3D protein structures is  
782 growing each year. In principal the docking score technology and statistical  
783 analyses methodology we present can scale to large numbers, the actual scaling  
784 behavior has yet to be characterized. As new proteins and new drugs are added  
785 to our calculations, we would expect quadratic scaling in the drug x protein  
786 matrix. The machine learning algorithms used to learn statistical correlations from  
787 this data should scale as a higher-degree polynomial of the number of training  
788 samples, i.e. docking profile of a drug. The benefit of utilizing an HPC platform is  
789 that the effects of non-linear scaling can be addressed by the allocation of  
790 additional compute nodes and processors. Investigation of the actual scaling  
791 behavior with increasing data set size and increasing number of CPUs remains  
792 to be done as future work.

793 For ADR phenotypes, we are currently limited by the availability of clinical  
794 data on ADR phenotypes linked to drugs and publicly available ADR outcomes

795 data will always be biased toward approved drugs. Our results point to the  
796 importance of the target proteins, which might not be known for nascent  
797 compounds. The known or intended targets appear to be important for ADRs  
798 associated with major organ systems (e.g. renal, hepatic, and cardiac). Results  
799 of the toxicity panel analysis indicate that even at the MM/GBSA level, we need  
800 to improve the drug-target interaction estimates, as shown by the poor  
801 performing median AUCs for the ADR groups, endocrine, psychiatric, and renal.  
802 Everything else seems to track well with two layers of noise: drug-protein binding  
803 as an indicator of multi-factorial diseases and docking scores as indicators. Also,  
804 the minimal "comprehensive" set of proteins needed to obtain high-quality ADR  
805 prediction models is unknown. As more proteins and pathways are associated  
806 with ADR phenotypes, the minimal comprehensive set will be soon be obtained.

807       There are also limitations associated with the way we corroborated  
808 putative ADR-protein associations with literature studies. Biological terms are  
809 used ambiguously in the literature. It was not our intent to find and report  
810 accurate, exhaustive numbers of papers that contained a particular putative  
811 ADR-protein correlation in PubMed. Rather, we wanted a well-defined, (i.e.  
812 UniProt names for proteins and MedDRA lowest-level terms for side effects)  
813 standardized way that would allow us to see a preponderance (e.g. more than  
814 10) of papers in the literature, where we could then obtain a sample and examine  
815 the quality of the correlation manually. Any other approach (e.g. stemming the  
816 terms) would have some ambiguity associated with it.

817

## 818 **CONCLUSIONS**

819 We have shown in this study that molecular docking may provide reliable, cost-  
820 effective, comprehensive, high-throughput screening of a drug candidate for  
821 binding across many known targets to provide predictions of clinically important  
822 ADRs. By introducing a first principles approach to *in silico* ADR prediction for  
823 drug compounds that heavily leverages physics-based models and HPC, we  
824 docked 560 small molecule drugs to 409 structures of identified DrugBank  
825 protein targets. Only 21% (87 out of 409) of the drug-protein binding features  
826 involve known targets of the drug subset, providing a significant probe of off-  
827 target effects. The median AUCs obtained during 10-fold cross-validation were  
828 comparable between the VinaLC off-target models (AUC=0.60-0.69), and the  
829 DrugBank on-target models (AUC=0.61-0.74) across the ten ADR groups. Most  
830 importantly, the VinaLC off target model out performed the DrugBank on target  
831 model for predicting two ADR group, neoplasms and vascularDisorders. We  
832 further investigated the associations between the ten ADR groups and a  
833 consensus subset of 16 proteins used in early-stage in vitro toxicity screening  
834 panels. The analysis identified several putative ADR-protein associations.  
835 Successful PubMed queries found published results in support of these putative  
836 ADR-protein associations. For example, several associations between  
837 neoplasm-related ADRs and known tumor suppressor (Syk) and tumor  
838 invasiveness marker (MMP-1 and C3) proteins are found. Many of these  
839 associations involve off-target proteins and would not have been found using  
840 only the available drug-target data. Thus, increasing the reliability of the drug-

841 protein binding calculations and increasing the protein target set to include more  
842 proteins outside the known protein targets in DrugBank should identify additional  
843 off-target proteins which are associated with possible ADRs. This predictive  
844 computational platform would be advantageous during the drug development  
845 stage to predict ADRs of drug candidates such that candidates could be dropped  
846 or redesigned at an earlier stage.

847

848

#### 849 **Acknowledgements**

850 Funding was provided by Laboratory Directed Research and Development  
851 (LDRD) (004-SI-012). We also thank Livermore Computing Grand Challenge for  
852 extensive computing resources. This work was performed under the auspices of  
853 the U.S. Department of Energy by Lawrence Livermore National Laboratory  
854 under Contract DE-AC52-07NA27344. LLNL-JRNL-652054.

855

856

857

858 **REFERENCES**

859

860 [1] Giacomini KM, Krauss RM, Roden DM, Eichelbaum M, Hayden MR, et al.

861 (2007) When good drugs go bad. *Nature* 446: 975-977.

862 [2] Ernst FR and Grizzle AJ (2001) Drug-related morbidity and mortality: updating

863 the cost-of-illness model. *J Am Pharm Assoc (Wash)* 41: 192-199.

864 [3] Herman WH (2013) The Economic Costs of Diabetes: Is It Time for a New

865 Treatment Paradigm? *Diabetes Care* 36: 775-776.

866 [4] Komajda M, McMurray JJ, Beck-Nielsen H, Gomis R, Hanefeld M, et al.

867 (2010) Heart failure events with rosiglitazone in type 2 diabetes: data from the

868 RECORD clinical trial. *Eur Heart J* 31: 824-831.

869 [5] Baron JA, Sandler RS, Bresalier RS, Lanas A, Morton DG, et al. (2008)

870 Cardiovascular events associated with rofecoxib: final analysis of the APPROVe

871 trial. *Lancet* 372: 1756-1764.

872 [6] Bowes J, Brown AJ, Hamon J, Jarolimek W, Sridhar A, et al. (2012) Reducing

873 safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev*

874 *Drug Discov* 11: 909-922.

875 [7] Whitebread S, Hamon J, Bojanic D and Urban L (2005) Keynote review: In

876 vitro safety pharmacology profiling: an essential tool for successful drug

877 development. *Drug Discovery Today* 10: 1421-1433.

- 878 [8] Fliri AF, Loging WT, Thadeio PF and Volkmann RA (2005) Analysis of drug-  
879 induced effect patterns to link structure and side effects of medicines. *Nat Chem*  
880 *Biol* 1: 389-397.
- 881 [9] Cobanoglu MC, Liu C, Hu F, Oltvai ZN and Bahar I (2013) Predicting Drug-  
882 Target Interactions Using Probabilistic Matrix Factorization. *Journal of Chemical*  
883 *Information and Modeling* 53: 3399-3409.
- 884 [10] Campillos M, Kuhn M, Gavin A-C, Jensen LJ and Bork P (2008) Drug Target  
885 Identification Using Side-Effect Similarity. *Science* 321: 263-266.
- 886 [11] Mizutani S, Pauwels E, Stoven V, Goto S and Yamanishi Y (2012) Relating  
887 drug-protein interaction network with drug side effects. *Bioinformatics* 28: i522-  
888 i528.
- 889 [12] Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, et al. (2008)  
890 SuperTarget and Matador: resources for exploring drug-target relationships.  
891 *Nucleic Acids Res* 36: D919-922.
- 892 [13] Kuhn M, Campillos M, Letunic I, Jensen LJ and Bork P (2010) A side effect  
893 resource to capture phenotypic effects of drugs. *Mol Syst Biol* 6: 343.
- 894 [14] Kanehisa M (2002) The KEGG database. *Novartis Found Symp* 247: 91-  
895 101; discussion 101-103, 119-128, 244-152.
- 896 [15] Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene  
897 Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32: D258-  
898 261.
- 899 [16] Pauwels E, Stoven V and Yamanishi Y (2011) Predicting drug side-effect  
900 profiles: a chemical fragment-based approach. *BMC Bioinformatics* 12: 169.

901 [17] Bolton EE, Wang Y, Thiessen PA and Bryant SH (2008) Chapter 12  
902 PubChem: Integrated Platform of Small Molecules and Biological Activities. In: A.  
903 W. Ralph and C. S. David, editors. Annual Reports in Computational Chemistry.  
904 Elsevier. pp. 217-241.

905 [18] Yamanishi Y, Pauwels E and Kotera M (2012) Drug side-effect prediction  
906 based on the integration of chemical and biological spaces. J Chem Inf Model 52:  
907 3284-3292.

908 [19] Liu M, Wu Y, Chen Y, Sun J, Zhao Z, et al. (2012) Large-scale prediction of  
909 adverse drug reactions using chemical, biological, and phenotypic properties of  
910 drugs. Journal of the American Medical Informatics Association 19: e28-e35.

911 [20] Cami A, Arnold A, Manzi S and Reis B (2011) Predicting Adverse Drug  
912 Events Using Pharmacological Network Models. Science Translational Medicine  
913 3: 114ra127.

914 [21] Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, et al.  
915 (2013) Systematic identification of proteins that elicit drug side effects. Mol Syst  
916 Biol 9: 663.

917 [22] Tatonetti NP, Liu T and Altman RB (2009) Predicting drug side-effects by  
918 chemical systems biology. Genome Biol 10: 238.

919 [23] Scheiber J, Chen B, Milik M, Sukuru SCK, Bender A, et al. (2009) Gaining  
920 Insight into Off-Target Mediated Effects of Drug Candidates with a  
921 Comprehensive Systems Chemical Biology Analysis. Journal of Chemical  
922 Information and Modeling 49: 308-317.

923 [24] Huang LC, Wu X and Chen JY (2011) Predicting adverse side effects of  
924 drugs. *BMC Genomics* 12 Suppl 5: S11.

925 [25] Huang LC, Wu X and Chen JY (2013) Predicting adverse drug reaction  
926 profiles by integrating protein interaction networks with drug structures.  
927 *Proteomics* 13: 313-324.

928 [26] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The  
929 Protein Data Bank. *Nucleic Acids Res* 28: 235-242.

930 [27] Yang L, Chen J and He L (2009) Harvesting candidate genes responsible for  
931 serious adverse drug reactions from a chemical-protein interactome. *PLoS*  
932 *Comput Biol* 5: e1000441.

933 [28] Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, et al. (2012)  
934 Large-scale prediction and testing of drug activity on side-effect targets. *Nature*  
935 486: 361-367.

936 [29] Wallach I, Jaitly N and Lilien R (2010) A structure-based approach for  
937 mapping adverse drug reactions to the perturbation of underlying biological  
938 pathways. *PLoS One* 5: e12063.

939 [30] Reardon S (2013) Project ranks billions of drug interactions. *Nature* 503:  
940 449-450.

941 [31] Kirkpatrick P and Ellis C (2004) Chemical space. *Nature* 432: 823-823.

942 [32] Zhang X, Wong SE and Lightstone FC (2013) Message passing interface  
943 and multithreading hybrid for parallel molecular docking of large databases on  
944 petascale high performance computing machines. *J Comput Chem* 34: 915-927.

945 [33] Zhang X, Wong SE and Lightstone FC (2014) Toward Fully Automated High  
946 Performance Computing Drug Discovery: A Massively Parallel Virtual Screening  
947 Pipeline for Docking and Molecular Mechanics/Generalized Born Surface Area  
948 Rescoring to Improve Enrichment. J Chem Inf Model.

949 [34] FDA (2014) Orange Book: Approved Drug Products with Therapeutic  
950 Equivalence Evaluations.

951 [35] <http://www.accessdata.fda.gov/scripts/Cder/ob/default.cfm>: FDA.

952 Brown EG, Wood L and Wood S (1999) The medical dictionary for regulatory  
953 activities (MedDRA). Drug Saf 20: 109-117.

954 [36] Nilmeier JP, Kirshner DA, Wong SE and Lightstone FC (2013) Rapid  
955 catalytic template searching as an enzyme function prediction procedure. PLoS  
956 One 8: e62535.

957 [37] Kirshner DA, Nilmeier JP and Lightstone FC (2013) Catalytic site  
958 identification--a web server to identify catalytic site structural matches throughout  
959 PDB. Nucleic Acids Res 41: W256-265.

960 [38] Halgren TA (2009) Identifying and Characterizing Binding Sites and  
961 Assessing Druggability. Journal of Chemical Information and Modeling 49: 377-  
962 389.

963 [39] Richards FM (1977) Areas, volumes, packing and protein structure. Annu  
964 Rev Biophys Bioeng 6: 151-176.

965 [40] Kuntz ID, Blaney JM, Oatley SJ, Langridge R and Ferrin TE (1982) A  
966 geometric approach to macromolecule-ligand interactions. J Mol Biol 161: 269-  
967 288.

968 [41] Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, et al. (2005) The  
969 Amber biomolecular simulation programs. *J Comput Chem* 26: 1668-1688.

970 [42] Onufriev A, Bashford D and Case DA (2004) Exploring protein native states  
971 and large-scale conformational changes with a modified generalized born model.  
972 *Proteins: Structure, Function, and Bioinformatics* 55: 383-394.

973 [43] Zhang X, Gibbs AC, Reynolds CH, Peters MB and Westerhoff LM (2010)  
974 Quantum mechanical pairwise decomposition analysis of protein kinase B  
975 inhibitors: validating a new tool for guiding drug design. *J Chem Inf Model* 50:  
976 651-661.

977 [44] Thompson DC, Humblet C and Joseph-McCarthy D (2008) Investigation of  
978 MM-PBSA rescoring of docking poses. *J Chem Inf Model* 48: 1081-1091.

979 [45] Guimaraes CR and Cardozo M (2008) MM-GB/SA rescoring of docking  
980 poses in structure-based lead optimization. *J Chem Inf Model* 48: 958-970.

981 [46] Rastelli G, Rio AD, Degliesposti G and Sgobba M (2010) Fast and accurate  
982 predictions of binding free energies using MM-PBSA and MM-GBSA. *Journal of*  
983 *Computational Chemistry* 31: 797-810.

984 [47] Moitessier N, Englebienne P, Lee D, Lawandi J and Corbeil CR (2008)  
985 Towards the development of universal, fast and highly accurate docking/scoring  
986 methods: a long way to go. *Br J Pharmacol* 153 Suppl 1: S7-26.

987 [48] Raha K, Peters MB, Wang B, Yu N, Wollacott AM, et al. (2007) The role of  
988 quantum mechanics in structure-based drug design. *Drug Discov Today* 12: 725-  
989 731.

990 [49] Wang R, Lu Y and Wang S (2003) Comparative Evaluation of 11 Scoring  
991 Functions for Molecular Docking. *Journal of Medicinal Chemistry* 46: 2287-2303.

992 [50] Sousa SF, Fernandes PA and Ramos MJ (2006) Protein-ligand docking:  
993 current status and future challenges. *Proteins* 65: 15-26.

994 [51] Yang Y, Lightstone FC and Wong SE (2013) Approaches to efficiently  
995 estimate solvation and explicit water energetics in ligand binding: the use of  
996 WaterMap. *Expert Opinion on Drug Discovery* 8: 277-287.

997 [52] Wong SE and Lightstone FC (2011) Accounting for water molecules in drug  
998 design. *Expert Opin Drug Discov* 6: 65-74.

999 [53] Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso.  
1000 *Journal of the Royal Statistical Society Series B* 58: 267-288.

1001 [54] Wu TT, Chen YF, Hastie T, Sobel E and Lange K (2009) Genome-wide  
1002 association analysis by lasso penalized logistic regression. *Bioinformatics* 25:  
1003 714-721.

1004 [55] Friedman J, Hastie T and Tibshirani R (2010) Regularization Paths for  
1005 Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33: 1-22.

1006 [56] Storey JD and Tibshirani R (2003) Statistical significance for genomewide  
1007 studies. *Proc Natl Acad Sci U S A* 100: 9440-9445.

1008 [57] Bonferroni C (1935) Il calcolo delle assicurazioni su gruppi di teste. *Studi in*  
1009 *Onore del Professore Salvatore Ortu Carboni*. Rome, Italy. pp. 13-60.

1010 [58] Chimal-Ramirez GK, Espinoza-Sanchez, NA et al. (2013) MMP1, MMP9,  
1011 and COX2 Expressions in Promonocytes Are Induced by Breast Cancer Cells  
1012 and Correlate with Collagen Degradation, Transformation-Like Morphological

1013 Changes in MCF-10A Acini, and Tumor Aggressiveness. *BioMed Research*  
1014 *International* 2013: 15.

1015 [59] Li X and Tai H-H (2013) Thromboxane A2 receptor-mediated release of  
1016 matrix metalloproteinase-1 (MMP-1) induces expression of monocyte  
1017 chemoattractant protein-1 (MCP-1) by activation of protease-activated receptor 2  
1018 (PAR2) in A549 human lung adenocarcinoma cells. *Molecular Carcinogenesis*:  
1019 doi: 10.1002/mc.22020.

1020 [60] Lin Y, Wang JF, Gao GZ, Zhang GZ, Wang FL, et al. (2013) Plasma levels  
1021 of tissue inhibitor of matrix metalloproteinase-1 correlate with diagnosis and  
1022 prognosis of glioma patients. *Chin Med J (Engl)* 126: 4295-4300.

1023 [61] Korematsu S, Ohta Y, Tamai N, Takeguchi M, Goto C, et al. (2012) Cell  
1024 distribution differences of matrix metalloproteinase-9 and tissue inhibitor of matrix  
1025 metalloproteinase-1 in patients with Kawasaki disease. *Pediatr Infect Dis J* 31:  
1026 973-974.

1027 [62] Chen XL, Li L and Zhang YJ (2011) [Regulatory role of Syk gene on vascular  
1028 endothelial growth factor C expression in breast cancer]. *Zhonghua Bing Li Xue*  
1029 *Za Zhi* 40: 805-809.

1030 [63] Zhang X, Shrikhande U, Alicie BM, Zhou Q and Geahlen RL (2009) Role of  
1031 the protein tyrosine kinase Syk in regulating cell-cell adhesion and motility in  
1032 breast cancer cells. *Mol Cancer Res* 7: 634-644.

1033 [64] Repana K, Papazisis K, Foukas P, Valeri R, Kortsaris A, et al. (2006)  
1034 Expression of Syk in invasive breast cancer: correlation to proliferation and  
1035 invasiveness. *Anticancer Res* 26: 4949-4954.

1036 [65] Layton T, Stalens C, Gunderson F, Goodison S and Silletti S (2009) Syk  
1037 Tyrosine Kinase Acts as a Pancreatic Adenocarcinoma Tumor Suppressor by  
1038 Regulating Cellular Growth and Invasion. *The American Journal of Pathology*  
1039 175: 2625-2636.

1040 [66] Dowling P, Clarke C, Hennessy K, Torralbo-Lopez B, Ballot J, et al. (2012)  
1041 Analysis of acute-phase proteins, AHSG, C3, CLI, HP and SAA, reveals  
1042 distinctive expression patterns associated with breast, colorectal and lung  
1043 cancer. *Int J Cancer* 131: 911-923.

1044 [67] Wilgenhof S, Morlion V, Seghers AC, Du Four S, Vanderlinden E, et al.  
1045 (2012) Sarcoidosis in a patient with metastatic melanoma sequentially treated  
1046 with anti-CTLA-4 monoclonal antibody and selective BRAF inhibitor. *Anticancer*  
1047 *Res* 32: 1355-1359.

1048 [68] Hattori N, Niimi T, Sato S, Achiwa H, Maeda H, et al. (2005) Cytotoxic T-  
1049 lymphocyte antigen 4 gene polymorphisms in sarcoidosis patients. *Sarcoidosis*  
1050 *Vasc Diffuse Lung Dis* 22: 27-32.

1051 [69] Romeo G, Frangioni JV and Kazlauskas A (2004) Profilin acts downstream  
1052 of LDL to mediate diabetic endothelial cell dysfunction. *FASEB J* 18: 725-727.

1053 [70] Stuijver DJ, van Zaane B, Romualdi E, Brandjes DP, Gerdes VE, et al.  
1054 (2012) The effect of hyperthyroidism on procoagulant, anticoagulant and  
1055 fibrinolytic factors: a systematic review and meta-analysis. *Thromb Haemost* 108:  
1056 1077-1088.

1057 [71] Kim HW, Rapoport SI and Rao JS (2010) Altered expression of apoptotic  
1058 factors and synaptic markers in postmortem brain from bipolar disorder patients.  
1059 Neurobiol Dis 37: 596-603.

1060 [72] Gasso P, Mas S, Molina O, Lafuente A, Bernardo M, et al. (2014) Increased  
1061 susceptibility to apoptosis in cultured fibroblasts from antipsychotic-naive first-  
1062 episode schizophrenia patients. J Psychiatr Res 48: 94-101.

1063 [73] Konstandin MH, Aksoy H, Wabnitz GH, Volz C, Erbel C, et al. (2009) Beta2-  
1064 integrin activation on T cell subsets is an independent prognostic factor in  
1065 unstable angina pectoris. Basic Res Cardiol 104: 341-351.

1066 [74] Pere JJ and Chaumet-Riffaud D (1990) [Clozapine and resistant  
1067 schizophrenia]. Encephale 16: 143-145.

1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

1080 **Figure Legends**

1081

1082 **Figure 1. Data integration/analysis workflow scheme.** The UniProt IDs of  
1083 4,020 proteins identified in DrugBank as drug targets were extracted. We  
1084 obtained 409 experimental protein structures from the Protein Data Bank (PDB)  
1085 to be used as a virtual panel and docked to 906 FDA-approved small molecule  
1086 compounds using the VinaLC docking code, run on a high-performance  
1087 computing machine at LLNL. 560 compounds had side effect information in the  
1088 SIDER database and were used in subsequent statistical analysis to build logistic  
1089 regression models for ADR prediction.

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102 **Figure 2. ADR prediction models using ‘Vina Off Targets’ and ‘DrugBank**  
1103 **On-Targets’.** Boxplots of median AUC results for one vs. all L1-regularized  
1104 logistic regression models trained using 10-fold cross-validation repeated ten  
1105 times are shown. The individual models were trained on ten different adverse  
1106 drug reaction (ADR) groups: Neoplasms, benign, malignant, and unspecified (N),  
1107 Blood and lymphatic systems disorders (B), Immune system disorders (I),  
1108 endocrine disorders (E), Psychiatric disorders (P), Cardiac disorders (C),  
1109 Vascular disorders (V), Gastrointestinal disorders (G), Hepatobiliary disorders  
1110 (H), and Renal Disorders (R). Red boxes indicate models trained on 560 x 409  
1111 VinaLC docking scores used as drug-protein binding features (also indicated by  
1112 .VD in the x-axis labels). Blue boxes indicate models trained on a 560 x 555  
1113 matrix containing DrugBank drug-target protein associations (also indicated by  
1114 the .DB appending the labels).

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124 **Figure 3. ADR prediction using a 16-protein virtual toxicity screening panel**  
1125 **suggested by Bowes et al.[6].** Red boxes (and the .G on the labels) indicate  
1126 models trained on GBSA-corrected VinalLC docking scores while the blue boxes  
1127 (and the .DB) indicate models trained on DrugBank drug-target protein  
1128 associations. The boxplots comprise the distribution of median AUC scores after  
1129 one vs. all L1-regularized logistic regression model training using 10-fold cross-  
1130 validation repeated ten times. The individual models were trained on ten different  
1131 adverse drug reaction (ADR) groups: Neoplasms, benign, malignant, and  
1132 unspecified (N), Blood and lymphatic systems disorders (B), Immune system  
1133 disorders (I), endocrine disorders (E), Psychiatric disorders (P), Cardiac  
1134 disorders (C), Vascular disorders (V), Gastrointestinal disorders (G),  
1135 Hepatobiliary disorders (H), and Renal Disorders (R).

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147 Table I. Top-ranked ADR-protein associations derived from models built using  
 1148 the 560 x 409 docking score matrix. The docked protein responsible for the  
 1149 association with the ADR is identified in the first, second columns, and third  
 1150 columns using the UniProt name and ID and the corresponding PDB ID,  
 1151 respectively. Columns 4,5,6 give data on the statistical significance of the  
 1152 association with the p-value of the association, the associated false discovery  
 1153 rate (q-value), and the corresponding beta coefficient in the median AUC logistic  
 1154 regression model. Bold UniProt IDs are off-target proteins (i.e. not intended  
 1155 targets of the 732 drugs we consider).  
 1156

UniProtName	UniProtID	PDBID#	p-value	q-value	beta	UniProtprotein-MedDRAsideEffectPubMedhits
InterstitialCollagenase	<b>P03956</b>	1hfc	0.004	0.531	2.348	breastneoplasm(158),denocarcinoma(161),glioma(34),basalcellcarcinoma(22)
Tyrosine-proteinKinaseTYK	<b>P43405</b>	1xbb	0.012	0.531	1.213	breastneoplasm(46),denocarcinoma(11)
Peroxisomeproliferator-activatedreceptorAlpha	Q07869	2znn	0.016	0.531	0.602	breastneoplasm(95),denocarcinoma(146),glioma(25),basalcellcarcinoma(14)
ComplementC3	<b>P01024</b>	2wy8	0.034	0.531	0.698	breastneoplasm(65),denocarcinoma(136),glioma(21),lungneoplasmsmalignant(12),basalcellcarcinoma(7)
CytotoxicT-lymphocyteprotein	<b>P16410</b>	3osk	0.003	0.555	0.211	sarcoidosis(11),vasculitis(24)
Profilin-1	<b>P07737</b>	1fil	0.000	0.005	0.338	endocrine disorder(10)
CoagulationFactorX	<b>P00740</b>	1edm	0.000	0.005	0.019	endocrine disorder(108),diabetesmellitus(48),thyroid disorder(22),hyperthyroidism(11),hypothyroidism(10)
Interleukin-5	<b>P05113</b>	1hul	0.000	0.005	0.092	endocrine disorder(35),diabetesmellitus(19),thyroid disorder(10)
Caspase-3	<b>P42574</b>	2dko	0.002	0.188	-1.876	bipolar disorder(14),schizophrenia(31)
IntegrinBeta-2	P05107	2p26	0.020	1.000	-0.886	cardiac arrest(11),cardiomyopathy(44),myocardial infarction(46)
InterstitialCollagenase	<b>P03956</b>	1hfc	0.000	0.060	0.429	aneurysm(39),orticaneurysm(31),arteriosclerosis(123)
Gelsolin	<b>P06396</b>	2fh1	0.000	0.009	-0.073	nephropathy(38),renal failure(12)

1157  
1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171 Table II. ADR-protein association derived from models built using the 560 x 16

1172 GBSA-corrected virtual screening panel.

1173

UniProt Name	UniProt ID	Corrected p-value	ADR Group	UniProt protein MedDRA side effect PubMed hits
Amine oxidase [flavin-containing] A	P21397	0.005	bloodAndLymph	agranulocytosis(5)
Histamine H1 receptor	P35367	0.007	bloodAndLymph	agranulocytosis(10)
Beta-2 adrenergic receptor	P07550	0.007	endocrineDisorders	endocrine disorder(164), diabetes mellitus(98), thyroid disorder(31), hyperthyroidism(19), hypothyroidism(16)
5-hydroxytryptamine receptor 1B	P28222	0.007	endocrineDisorders	endocrine disorder(15), diabetes mellitus(11)
Androgen receptor	P10275	0.018	psychDisorders	schizophrenia(18)
Prostaglandin G/H synthase 2	P35354	0.024	cardiacDisorders	cardiac arrest(11), cardiomegaly(22), cardiomyopathy(91), myocardial infarction(217), myocarditis(11)

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192 **Supplementary Information Titles and Files**

1193

1194 Supplementary Figure 1. Protein structure quality control workflow used in  
1195 preparation for docking calculations

1196 <File name: Supplementary\_Figure1.eps, File type: EPS file>

1197

1198 Supplementary Table 1. ADR groupings and their MedDRA LLT side effect  
1199 components

1200 <File name: Supplementary\_Table1.xlsx, File type: MS Excel Workbook file>