Peer-reviewed version available at Molecules 2019, 24, 645; doi:10.3390/molecules24030645

*Article*

# Small molecule docking of DNA repair proteins associated with cancer survival following PCNA metagene adjustment: A potential novel class of repair inhibitors

**Leif E. Peterson**[1,2]

[1] Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York City, New York 10065 USA
[2] Center for Biostatistics, Institute for Academic Medicine, Houston Methodist Research Institute, 6565 Fannin Street, Houston, Texas 77030 USA
* Correspondence: peterson.leif.e@gmail.com; Tel.: +001 (281) 381-6218

**Abstract:** Natural and synthetic small molecules from the NCI Developmental Therapeutics Program (DTP) were employed in molecular dynamics-based docking with DNA repair proteins whose RNA-Seq based expression was associated with overall cancer survival (OS) after adjustment for the PCNA metagene. The compounds employed were required to elicit a sensitive response (vs. resistance) in more than half of the cell lines tested for each cancer. Methodological approaches included peptide sequence alignments and homology modeling for 3D protein structure determination, ligand preparation, docking, toxicity and ADME prediction. Docking was performed for unique lists of DNA repair proteins which predict OS for AML, cancers of the breast, lung, colon, and ovaries, GBM, melanoma, and renal papillary cancer. Results indicate hundreds of drug-like and lead-like ligands with best-pose binding energies less than -6 kcal/mol. Ligand solubility for the top 20 drug-like hits approached lower bounds, while lipophilicity was acceptable. Most ligands were also blood-brain barrier permeable with high intestinal absorption rates. While the majority of ligands lacked positive prediction for Herg channel blockage and Ames carcinogenicity, there was considerable variation for predicted fathead minnow, honey bee, and Tetrahymena pyriformis toxicity. The computational results suggest the potential for new targets and mechanisms of repair inhibition and can be directly employed for in vitro and in vivo confirmatory laboratory experiments to identify new targets of therapy for cancer survival.

**Keywords:** Small molecule; ligand; receptor; docking; molecular dynamics; DNA repair; inhibition, PCNA, ADME, toxicology.

## 1. Introduction

One of the hallmarks of cancer is cellular growth dysregulation caused by mutations as a result of genomic instability[1, 2]. Such mutations play an important role in oncogenic transformation and can be catastrophic during mitosis, or lead to chromothripsis[3, 4]. The continuous forced cell division in tumor cells also results in replication stress and increased oxidative damage, which requires several DNA repair components [5, 6]. When DNA repair deficiencies occur as a result of oncogenic loss or genetic polymorphisms, alternative DNA repair pathways must be found if replication is to continue[7]. Cancer's addiction to alternative DNA repair pathways can therefore be targeted to prevent the repair and restart of stressed replication forks [8-10].

Genetic stability relies on DNA repair, which is a complex process that depends on several molecular pathways to correct damage to DNA. DNA damage ranges from minor mismatched bases and methylation events to oxidized bases, intra- and interstrand DNA crosslinks, protein-DNA

adducts, double strand breaks (DSBs), and stalled forks. DNA repair pathways include mismatch repair, base excision repair, nucleotide excision repair, and the homology directed repair/Fanconi anemia pathway. Earlier forms of chemotherapy and radiation therapy focused on damaging DNA to promote excessive lethal mutations and cellular death; however, it has been demonstrated that cancer cells can repair therapy-induced DNA damage[11, 12]. This has led to the concept of synthetic lethality and DNA repair inhibition, in which specific DNA repair pathways and their proteins are targeted for increasing sensitivity to traditional therapeutics[13, 14].

DNA repair inhibitors have been introduced to augment molecular-based therapies for oncogene addiction and synthetic lethality [11, 12, 14]. DNA repair inhibitors for oncotherapy fall into several classes, which include poly (ADP) ribose polymerase (PARP) inhibitors, nucleotide excision repair (NER) inhibitors, DNA-PK inhibitors, MRN, ATM, ATR, CHK1/2 inhibitors, RAD51 inhibitors, and base excision repair (BER) inhibitors. PARP inhibitors have demonstrated great promise in the treatment of patients with deficiencies in homologous recombination (HR) DNA repair, such as those with loss of BRCA1 or BRCA2 function[11, 15-20]. Nucleotide excision repair inhibitors target more than thirty protein-protein interactions and removes DNA adducts caused by platinum-based chemotherapy[21-25]. DNA-PK inhibitors [26-30] target DNA-dependent protein kinase (DNA-PK) enzymes, which play a role in detection and repair of DSB via the non-homologous end-joining pathway. MRN, ATM, ATR, CHEK1/2 inhibitors [31, 32] target the kinases MRN, ATM, ATR, CHK1, and CHK2, and RAD51 [33-36] inhibitors target RAD51, a key protein of homologous recombination to repair DSB and inter-strand cross-links. BER inhibitors [37, 38] target BER proteins, which can protect a cell after endogenous or exogenous genotoxic stress, since a deficiency in BER can result in stress-induced apoptosis, necrotic cell death, mutagenesis, and chromosomal rearrangements.

In a recent investigation of TCGA RNA-Seq data and DNA repair genes, we identified sets of DNA repair genes for various cancers (Table 1) whose down-regulated expression patterns were associated with prolonged overall survival (OS)[39]. Prior to gene identification, DNA repair gene expression was adjusted for age at diagnosis, stage, and expression of the PCNA metagene[40]. Statistical randomization tests were also employed in which sets of DNA repair genes were randomly sampled for generating empirical, distribution-free, p-values. Using the list of DNA repair genes whose down-regulation was associated with longer OS, we hypothesized that compounds which strongly bind to these repair proteins could potentially establish new leads for novel DNA repair inhibitors. Additional insight could be gained from our use of the PCNA metagene to adjust expression of DNA repair genes prior to survival prediction, since this has heretofore eluded systematic investigation. Therefore, it warrants noting that the DNA repair genes in Table 1 would not have been identified without PCNA adjustment, and it is for this reason we believe this new perspective could very well define new targets for cancer therapy.

The purpose of this investigation was to perform in silico structural drug discovery hinged to molecular dynamics (MD) to identify which compounds from the US National Cancer Institute's (NCI) Developmental Therapeutics Program (DTP) repository tightly bind to DNA proteins identified. The DTP compounds used for ligand-receptor docking analysis will be filtered from the entire list of DTP-tested compounds, for compounds that are associated with cell line sensitivity rather than resistance. Computational toxicology and absorption, distribution, metabolism, and excretion (ADME) will also be employed to address potential safety concerns on a preliminary basis. This initial investigation can be followed by future in vitro and in vivo experiments such as qPCR and mouse/patient-derived xenografts to establish supportive lines of evidence for efficacy.

**Table 1.** DNA Repair proteins employed in ligand-receptor docking[39].

| Cancer | DNA repair proteins whose downregulation is associated with prolonged OS |
|---|---|
| AML | *RAD23A, EME2, APEX2* |
| Breast | *ATRIP, FANCC, RAD1, RFC3, NEIL3, EXO1, FANCB, FANCD2, FANCI, RAD51, XRCC4* |
| Colon | *RAD23A, RFC2, POLL, MLH3, FANCL* |
| GBM | *XRCC5, NBN, DDB1, GTF2H2, ERCC4, ALKBH2, APEX1, PRKDC, PMS1, REV1* |
| Renal papillary | *BLM, RAD1, FEN1, LIG1, EXO1, MSH6, BRCA2, EME1, FANCB, LIG4* |
| Lung | *BRCA1, NBN, RAD1, NEIL3, MMS19, FANCI, XRCC1, XRCC5* |
| Melanoma | *MDC1, NBN, MUTYH, POLE, UNG, FANCE, FANCI, POLI, POLK* |
| Ovarian | *PARP2, GTF2H4, SMUG1, DCLRE1B, GPS1, FANCL, APEX2, PMS1, XRCC6, MSH6, UNG, RAD51, RAD23A, EXO1, MUS81* |

## 2. Results

Table 2 lists the number of docked ligands as well as the number of drug-like and lead-like hits. The ligands that underwent docking were required to have a sensitive dose-response (z-score>0.5) in more than half of the DTP cell lines representing each cancer. Drug-like and lead-like criteria for physio-chemical properties were only applied to ligands whose receptor binding energies were <-6 kcal/mol for the best docking pose. For AML, among 3835 docked ligands, 1181 were drug-like and 173 were lead-like. For breast cancer, there were 684 ligands docked, for which 237 were drug-like and 47 were lead-like. GBM, on the other hand involved 343 docked ligands, for which 238 were drug-like and 38 were lead-like. Colon cancer ligand-receptor docking included 1123 ligands, with 751 being drug-like and 106 taking on lead-like properties. Lung cancer had the least number of docked ligands, with 31 identified having drug-activity, 20 that were drug-like, and 7 which were lead-like. Melanoma involved 291 docked ligands, with 224 revealing drug-like properties and 52 yielding lead-like properties. There were 105 docked ligands for ovarian cancer, with 75 drug-like and 8 lead-like. Finally, for renal papillary cancer there were 161 ligands docked, with 103 being drug-like, and 24 being lead-like.

**Table 2.** Number of docked ligands, and drug-like or lead-like ligands identified*.

| Cancer | Docked | Drug-like | Lead-like |
|---|---|---|---|
| AML | 3835 | 1181 | 173 |
| Breast | 684 | 237 | 47 |
| GBM | 343 | 238 | 38 |
| Colon | 1123 | 751 | 106 |
| Lung | 31 | 20 | 7 |
| Melanoma | 291 | 224 | 52 |
| Ovarian | 105 | 75 | 8 |
| Renal papillary | 161 | 103 | 24 |

*Docked compounds were required to have drug activity z-scores >0.5 in more than half of the DTP cell lines tested for each cancer. Docked candidates for drug-like and lead-like filtering were required to have binding energies <-6 kcal/mol for the best docking pose.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2019 doi:10.20944/preprints201901.0062.v1

Peer-reviewed version available at Molecules 2019, 24, 645; doi:10.3390/molecules24030645



For the cancers investigated, we were able to identify many energy-minimized compounds which were strongly bound to energy-minimized receptors. Altogether, our results indicate that many compounds were strongly bound to multiple receptors, and passed criteria for being drug-like, with fewer compounds portraying lead-like properties. The mean binding energy for ligand-receptor docking of each of the cancers considered was -7.42(1.22) for GBM, -7.19(1.2) for ovarian, -6.94(0.78) for lung, -6.93(1.25) for renal, -6.91(1.05) for melanoma, -6.88(1.04) for breast, -6.82(1.06) for colon, and -6.6(1.27) for AML. The top 10 most strongly bound receptors were PARP2(-8.88), REV1(-8.38), DDB1(-8.35), MUS81(-8.22), ALKB2(-8.01), XRCC5(-7.62), RFC3(-7.62), MUTYH(-7.5), POLI(-7.47), and FANCD2(-7.47), which revealed the importance of these receptors as potential druggable targets for therapy. Figure 1 shows the cancer-specific BE distribution for all possible ligand-receptor pairs. The large proportion of significant docking poses with BE<-6 kcal/mol are readily visible.

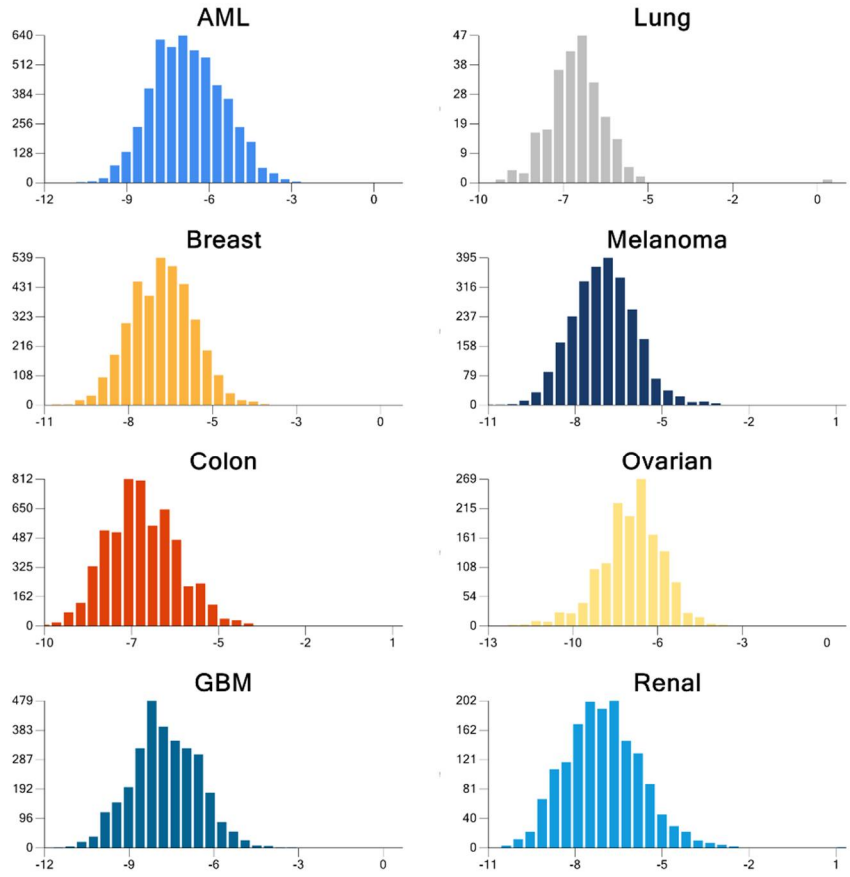



**Figure 1.** Distribution of the docking binding energy, BE (kcal/mol), from the best pose for all possible combinations of ligand-receptor pairs for each cancer. BE values <-6 kcal/mol are considered significant.

Figures 2-5 illustrate the 2D molecular structure for the top drug-like 20 ligands for AML, breast, colon, and lung cancer. Supplemental Figures S1-S4 contain 2D structure plots of the top 20 drug-like hits for GBM, melanoma, ovarian, and renal papillary cancer. Scaffold analysis for the drug-like and lead-like ligands followed by cluster analysis is now being pursued to determine if there are unique clusters of compounds. Tables 3-6 list the physio-chemical properties and computational ADME and toxicity predictions for the top 20 drug-like ligands shown in Figures 2-5. (Tables S1-S4 in Supplemental Information list these same parameters for GBM, melanoma, ovarian, and renal papillary cancers). As one can notice, lipophilicity values (LogP) fell within the acceptable range.

Compounds which are too lipophilic (LogP>5) are often associated with greater metabolic clearance, metabolite-related toxicity, lower solubility, less oral absorption, and affinity for hydrophobic targets instead of the desired target, which increases promiscuity-related off-target toxicity. Low lipophilicity can also increase renal clearance, and negatively impact permeability and decrease potency, resulting in lower efficacy. Most of the ligands reported also had solubilities (LogS) near lower bound of -4 for the majority of approved drugs; compounds with LogS<-6 are classified as being poorly soluble. Poorly soluble compounds tend to have poor absorption, low stability, and fast clearance[68]. Less soluble compounds are also more difficult to handle and formulate, and parenteral delivery requires greater levels of solubility[69]. Most of the top 20 drug-like ligands appeared to be BBB permeable and readily absorbed in the intestine (i.e., HIA) as indicated by high prediction probabilities. While Herg channel blockage and the Ames carcinogenicity tests did not seem to be of too much concern, there were several ligands which resulted in high probabilities for FHM, HBT, and TPT toxicity. However, during the stages of discovery, it is customary to sacrifice false positives (lower specificity) in toxicity, while prioritizing greater sensitivity for efficacy, due to the greater uncertainty in adverse events during clinical studies. There also appeared to be wide variation in the predicted inhibition of cytochrome P-450 (CYP) enzymes, which may or may not turn out to be a metabolic or safety concern. Our future in vitro and in vivo experiments will require additional filtering within the lists of drug-like and lead-like candidates (results not shown). In addition, further toxicity and ADME predictions will be pursued to refine these estimates.

## 3. Discussion

The adjustment of gene expression for the PCNA metagene in oncology is not a new concept[40]; however, there are groups which are unaware of its existence, and the novel perspective that can be attained once it is performed. Approximately 50% of the genome is co-regulated by PCNA, and because its effect is so widespread, it is imperative that investigations in cancer research account for and remove the effects of this non-cancer related transcriptional mechanism. Our effort to remove PCNA effects from RNA-Seq based gene expression in TCGA data is novel, and this has resulted in new lists of DNA repair genes which are predictive of OS beyond chance variation. Regarding molecular docking, virtual drug screens typically employ hundreds of thousands of ligands for which no dose-response information is available prior to the analysis. We took the opposite approach by filtering on cell line-derived drug activity z-scores, and only used DTP compounds which elicited a sensitive dose-response. Therefore, the only remaining uncertainties surrounding the strongly binding ligands are whether the observed DTP cell line sensitivities were due to a direct/indirect inhibitory mechanism, underlying toxicity independent of binding, or both. These issues can be addressed in future in vitro experiments to confirm protein binding and in vivo xenograft models to confirm animal model efficacy.

The results support repair inhibition models of cancer survival, which can be applied to longitudinal studies and clinical trials. The molecular docking employed revealed results that are potentially hinged to repair inhibition via protein binding. Overall, our docking analysis allowed us to make inferences about cell line sensitivity to draw conclusions about the relationships between binding and tumor growth for common cancers. The translational value of our results is established by the identification of drug-like and lead-like compounds. This could prove useful in future studies of molecular markers of therapy for delayed progression. Our future investigations will start with the top 50-100 ligand identified for each cancer to link verified gene-knockdown with xenograft models for confirming an association between repair inhibition and experimental data for tumor growth and survival. Molecular docking of DTP compounds with DNA repair proteins has enabled us to gain a perspective that hit identification and repair inhibition in the cell line data employed could likely reveal new compounds and mechanisms for oncotherapy. This view will hopefully enforce an appreciation among oncologists and biologists for the translational value of DTP cell line testing results, compounds, and molecular docking of these compounds with DNA repair proteins, for potential clinical trials involving single- or multi-label treatments associated with prolonged survival and pursuing longitudinal studies to improve therapeutic strategies.

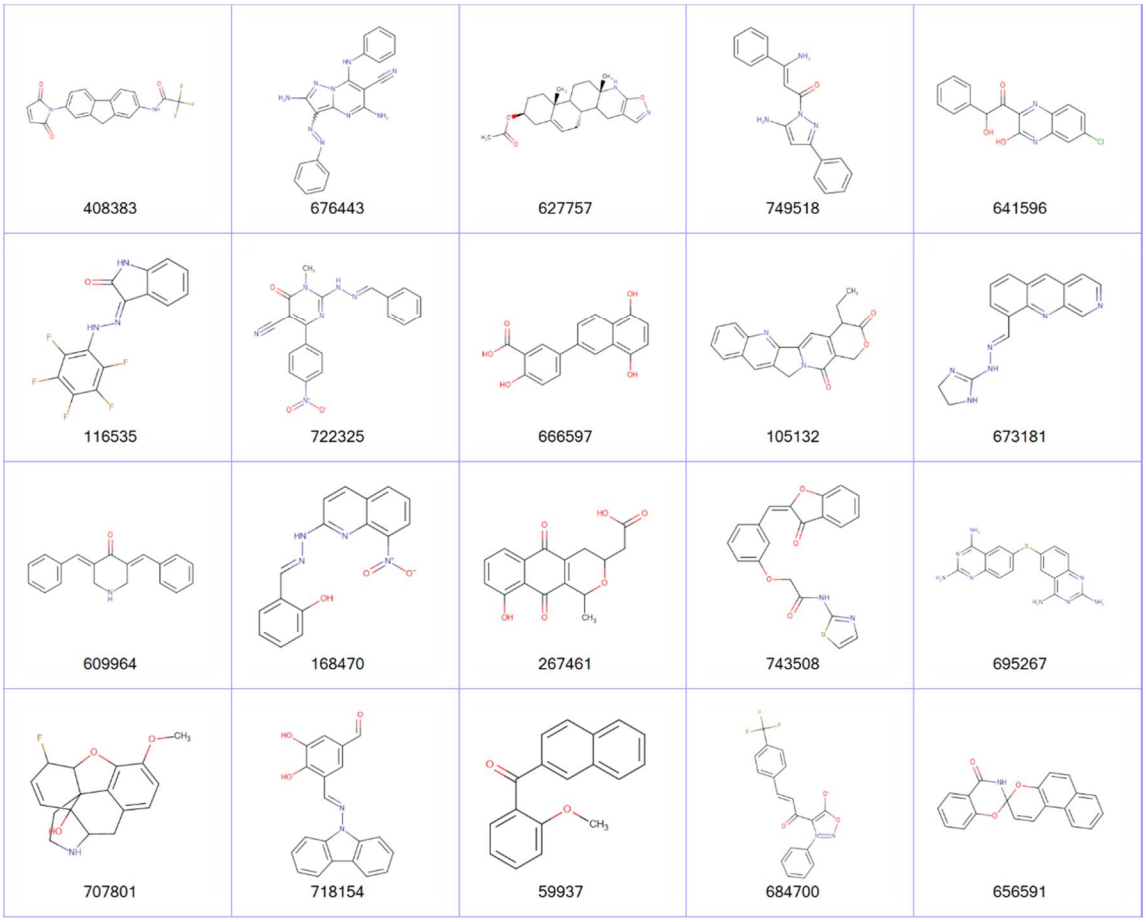



**Figure 2.** 2D structures of top 20 drug-like ligands for AML. Listed are NSC-408383, NSC-676443, NSC-627757, NSC-749518, NSC-641596, NSC-116535, NSC-722325, NSC-666597, NSC-105132, NSC-673181, NSC-609964, NSC-168470, NSC-267461, NSC-743508, NSC-695267, NSC-707801, NSC-718154, NSC-59937, NSC-684700, and NSC-656591.

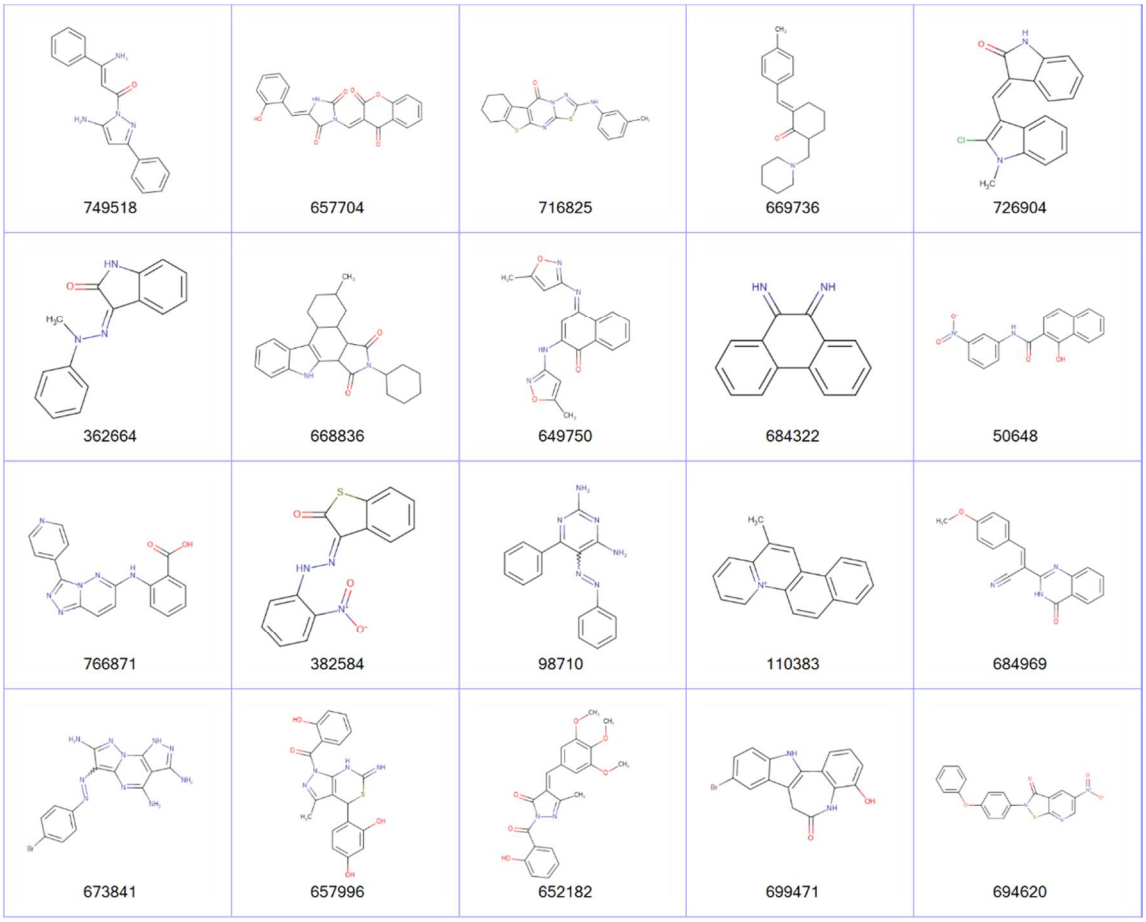
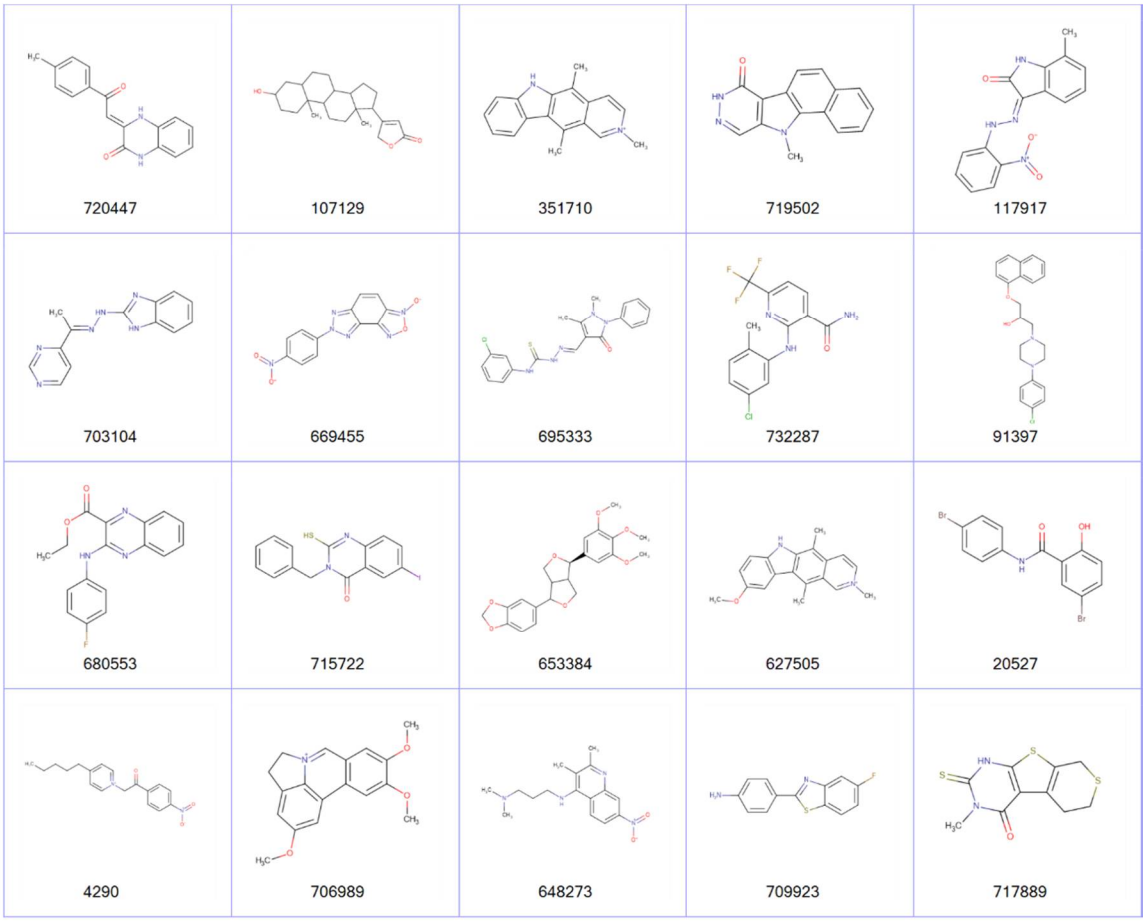



**Figure 3.** 2D structures of top 20 drug-like ligands for breast cancer. Listed are NSC-749518, NSC-657704, NSC-716825, NSC-669736, NSC-726904, NSC-362664, NSC-668836, NSC-649750, NSC-684322, NSC-50648, NSC-766871, NSC-382584, NSC-98710, NSC-110383, NSC-684969, NSC-673841, NSC-657996, NSC-652182, NSC-699471, and NSC-694620.

208




209

210 **Figure 4.** 2D structures of top 20 drug-like ligands for lung cancer. Listed are NSC-720447, NSC-
211 107129, NSC-351710, NSC-719502, NSC-117917, NSC-703104, NSC-669455, NSC-695333,NSC-732287,
212 NSC-91397, NSC-680553, NSC-715722, NSC-653384, NSC-627505, NSC-20527, NSC-4290, NSC-
213 706989, NSC-648273, NSC-709923, and NSC-717889.

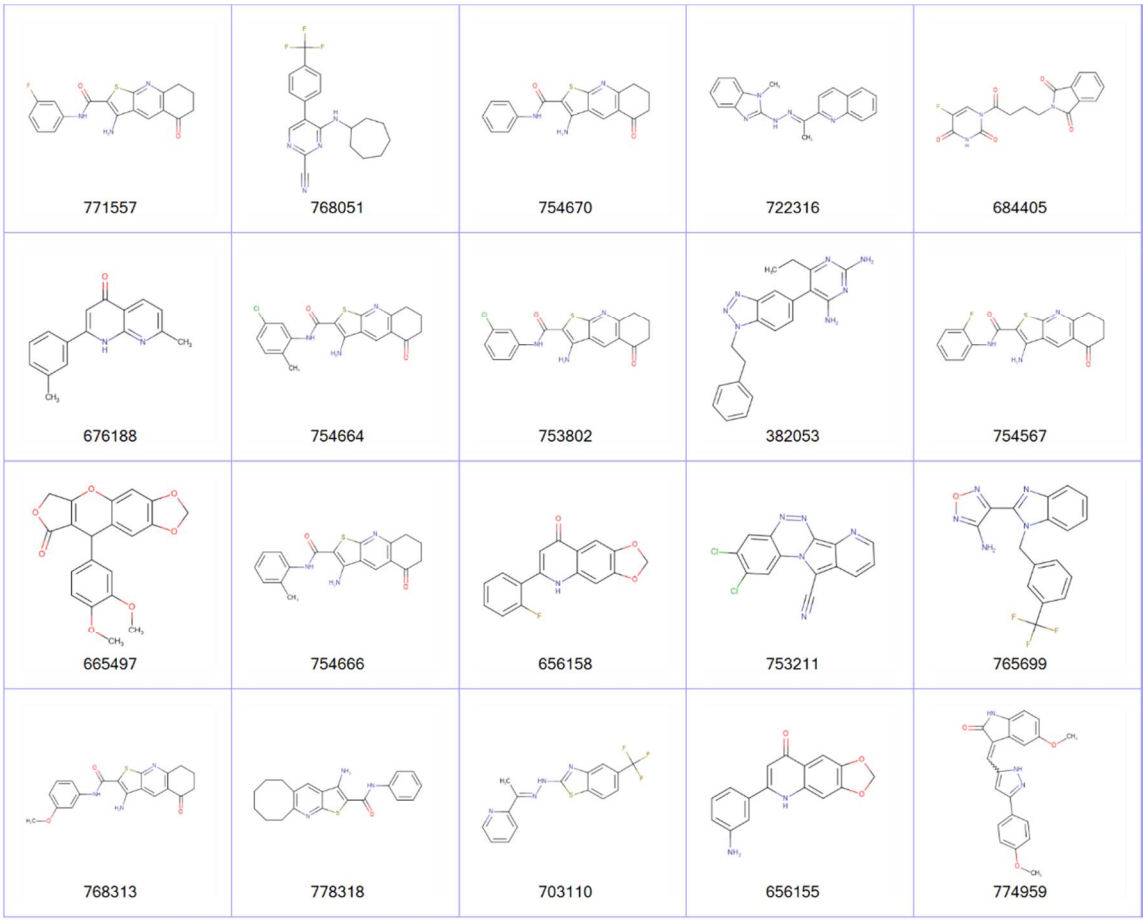



**Figure 5.** 2D structures of top 20 drug-like ligands for colon cancer. Listed are NSC-771557, NSC-768051, NSC-754670, NSC-722316, NSC-684405, NSC-676188, NSC-754664, NSC-753802, NSC-382053, NSC-754567, NSC-665497, NSC-754666, NSC-656158, NSC-753211, NSC-765699, NSC-768313, NSC-778318, NSC-703110, NSC-656155, and NSC-774959.

We did not comparatively assess gene expression in the tumor cell lines employed. Recall that although the DTP project has used a variety of gene expression platforms, expression patterns in cell lines was not available for most of the DNA repair genes used in this investigation. We also did not employ a gold-standard to establish false positive and false negative rates of toxicological and ADME endpoints. Rather, we ranked ligands by their binding energies hinged to MD and approximated computationally the toxicological and ADME outcomes. Experimental validation of the binding efficiency of ligands and in vitro determination of delayed tumor growth, toxicological, and ADME outcomes are needed in order to assess whether select groups of patients will benefit from new therapies arising from our findings. The work presented here suggests that molecular docking studies of DNA repair proteins with cell line-sensitive DTP compounds can provide new insights into the development of human cancer and can establish new leads for future research on molecular diagnosis and therapeutics.

There are several challenging issues surrounding development of tumor progression models based on repair inhibition. First, there is the problem of inherited germline polymorphisms, which confer a variety of risks and require a variety of treatment regimens. Second, tumor heterogeneity and immune escape are others hallmarks of cancer that cannot be easily overcome when searching for new modes of therapy. The RNA-Seq based data obtained from TCGA are not based on single-cell analysis, which would be helpful for elucidating heterogeneity; however, the large genetic variation identified throughout the TCGA samples used would exacerbate the complexity surrounding our attempt to portray repair inhibition via a single picture. Third, our approach was bioinformatic and not mechanistic using laboratory methods, since the intent was purely computational at this stage of investigation. We also did not consider DNA methylation status, chromosome aberrations, and microsatellite instability, which would overlay more complexity on the results obtained.

In conclusion, our bioinformatic approach to infer repair inhibition from cell line and RNA-Seq data in the absence of data on genomic alterations and germline polymorphisms should be cautiously interpreted. While most human cancers are nevertheless sporadic, the importance of the inherited component of cancer, combined with genomic alterations and tumor heterogeneity leading to growth and progression was not addressed in this investigation.

## 4. Materials and Methods

**Small-Molecule Ligand Library.** The NCI's DTP maintains a repository of synthetic and pure natural products that are available to investigators for non-clinical research purposes[41]. The repository collection is a uniquely diverse set of more than 200,000 compounds that have been either submitted to DTP for biological evaluation or, in some cases, synthesized under DTP auspices. Drug activity levels expressed as 50% growth inhibitory levels (GI50s) are determined by the DTP at 48 hours using the sulphorhodamine B assay[42]. All GI50 data was assessed and transformed to z-scores as described previously[43].

**Ligand Selection from DTP.** Activity z-scores for 21,737 DTP compounds tested for dose-response for each cell line were obtained using the CellMiner platform (V2.1 [44]). Activity z-scores >0.5 indicate cell line sensitivity, while z-scores <0.5 imply resistance. Compound activity z-scores were available for 60 cancer cell lines representing the following cancers: breast (5 cell lines), CNS(6), colon(7), leukemia(6), melanoma(10), lung(9), ovarian(7), prostate(2), and renal(8). For each cancer, a compound was selected if the z-score >0.5 for more than half of the cell lines available.

**Protein (receptor) 3D structures.** The receptors (proteins) employed in this investigation (Table 1) were identified during our previous work with TCGA RNA-Seq data and the PCNA metagene to have an association with prolonged overall survival (OS) when down-regulated. FASTA amino acid sequences (Homo Sapiens) for these proteins were obtained from Uniprot [45]. Homology modeling for obtaining a consensus 3D protein structure was determined using Swiss Model [46-48], which is based on Qmean [49], quaternary structure prediction/ QSQE[50], BLAST [51, 52], and the BLOSUM amino acid substitution matrix[53, 54], and were saved in PDB format.

Molecular charges were merged, and non-polar hydrogens, lone pairs, and water molecules were removed using the .NET assembly of OpenBabel (OB) [55].

**Molecular Ligand-Receptor Docking.** Ligand-receptor docking is an MD approach for reproducing chemical potentials which determine the bound conformation preference and free energy of binding between a ligand and its receptor. The MD technique seeks to establish the optimal receptor binding pocket (pose) with a minima in the energy profile, shape, and temperature [56], while assuming consistency in the ligand charge distribution and protonation states for the bound and unbound forms. At each receptor pocket identified, several poses are evaluated while iterating through alternative conformations of the ligand at its rotatable covalent bonds.

Prior to docking, ligand SMILES strings were converted to 3-dimensional SDF format containing partial charges of each atom. The .NET OB assembly was used for adding hydrogens to ligands and performing energy minimization of ligands and receptors using the Merck MMFF94 force field [57], with 250 iterations during conjugate gradient convergence. Energy-minimized ligands and receptors were saved in PDBQT format. Vina [58] was used for ligand-receptor docking on Amazon AWS cloud formations with Linux high-performance compute clusters. A total of 10 ligand poses were assessed at each receptor pocket identified, and the best pose was assumed to have the lowest binding energy (BE) in kcal/mol. BE values less than -6 kcal/mol are considered to represent significant binding affinity.

**Drug-like and Lead-like Hit Determination.** Ligands that yielded a best docking pose with BE<-6 were additionally filtered using physio-chemical properties of compounds. These included lipophilicity (LogP: log of octanal-water partition coefficient) and solubility (LogS) using the SMARTS notation available from SILICOS-IT[59], which were implemented in .NET. Molecular weight (MW), topological surface area (TPSA), number of hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and number of rotatable bonds (RotB) were determined using OB's .NET assembly. All compounds were kekulized and stripped of salts prior to calculation of physio-chemical properties, except for LogS solubility calculations, for which hydrogens were added. Two sets of criteria were employed for assessing suitability of ligands for lead discovery: "drug-like" and "lead-like". The drug-like hits were based on the Muegge (Bayer) criteria [60] for which $200 \leq MW \leq 600$, $-2 \leq LogP \leq 5$, $TPSA \leq 150$, $HBD \leq 5$, $HBA \leq 10$, and $RotB \leq 15$. Whereas the lead-like criteria were $LogP < 3$, $MW < 300$, $HBD \leq 3$, $HBA \leq 3$, and $RotB \leq 3$.

**Computational Toxicity and ADME Prediction.** Determination of the absorption, distribution, metabolism, excretion (ADME) and toxicity and of new and existing drugs is necessary to identify their harmful effects of humans, animals, plants, and their environment. Historically, in vivo animal models have been applied for ADME and toxicity testing; however, these are constrained by time, ethical considerations, and financial burden. As an alternative, in silico computational methods can be used to simulate, analyze, and visualize predictions for ADME and toxicity. In silico ADME and toxicology predictions can complement drug design to prioritize chemicals, guide toxicity tests, and minimize late-stage failure of new drugs. Computational prediction can also potentially minimize the need for animal testing, reduce costs and time for toxicity testing, and improve toxicity and safety assessment. Early-stage identification of hazardous new compounds can also improve the cost-benefits of novel drug synthesis.

**Fathead Minnow Toxicity (FMT).** The Fathead minnow is an important aquatic and terrestrial toxicity endpoint target, and Fathead minnow toxicity data were obtained from Cheng et al. [61]. FMT toxicity data consisted of 188 FMT- and 366 FMT+ compounds (554 total). The FMT endpoint for each compound was expressed as the concentration lethal to 50% of the organisms (LC50) for FMT during 96h flow-through exposure tests. Cheng et al. [61] selected a threshold value of LC50=0.5mmol/L to partition the data into low and high acute FMT compounds. Compounds with the value of LC50 less than 0.5 mmol/L were assigned as high acute FMT compounds, whereas others were assigned as low acute FMT compounds. The chemical name, CAS numbers, FMT test results, and SMILES strings were available in the data.

**Honey Bee Toxicity (HBT).** 195 pesticides or pesticide-like molecules for HBT (96 HBT-, 99 HBT+) were collected from Cheng et al. [61], based on data from the US EPA ECOTOX database[62].

The HBT endpoint for Apis mellifera bees was expressed as the dose lethal to 50% of the test population (LD50) during a 48h exposure test. Cheng et al. [61] selected a threshold value of LD50=100 μg/bee to designate high acute HBT compounds and low acute HBT compounds. Compounds with an LD50 below 100 μg/bee were coded as high acute HBT compounds, while others were coded as low acute HBT compounds. The chemical name, CAS numbers, HBT test results, and SMILES strings were available in the data.

**Tetrahymena Pyriformis Toxicity (TPT).** Tetrahymena pyriformis toxicity (TPT) is often used as a toxicology endpoint, and 1571 diverse TPT-tested chemicals were collected from Cheng et al. [63]. Toxicity data was expressed as the negative logarithm of 50% growth inhibitory concentration (pIGC50) values and duplicated molecules were removed. Xue et al. [64] selected a threshold value of pIGC50=-0.5 for discriminating TPT and non-TPT compounds (compounds with pIGC50>-0.5 were assigned as TPT, otherwise as non-TPT). The entire dataset was then divided into 1217 TPT+ and 354 TPT- compounds. The chemical name, CAS numbers, SMILES strings and pIGC50 value of 1571 compounds were available in the data.

**Human Intestinal Absorption (HIA).** The original HIA dataset was collected from Shen et al. [65]. This dataset contained n=578 compounds with fraction absorption (%FA) values. Shen et al. also specified a threshold value of 30% to partition compounds into HIA+ and HIA- (78 HIA- and 500 HIA+ compounds). Drugs with oral dosage formulations were considered to be HIA+ compounds. The chemical name, SMILES and class labels HIA+ and HIA- were available in the data.

**Blood Brain Barrier Penetration (BBB).** The BBB dataset contained n=1593 compounds, also obtained from Shen et al. [65], and have been categorized into BBB+ (n=1283) and BBB-(n=310). The chemical name, CAS numbers, BBB test results, and SMILES strings were available in the data.

**Cytochrome P450 Inhibition (CYP).** The P450 gene superfamily is involved in the metabolism of approximately 90% of approved drugs and clearance of xenobiotics. Drug safety and toxicity is directly hinged to CYP enzyme inhibition, because if a compound is a strong inhibitor of a CYP enzyme, it can result in reduced metabolism of drugs that are a substrate of the CYP enzyme, potentially leading to toxic serum plasma levels. Therefore, it is imperative to develop in silico prediction models of CYP inhibition for new compounds to project safety and toxicity. A large dataset containing more than 13,445 unique compounds against five major CYP isoforms, namely, 1A2, 2C9, 2C19, 2D6, and 3A4, was obtained from the PubChem AID-1851 database [66]. The assay employed for generation of these data used various human CYP P450 isozymes to measure the dealkylation of various pro-luciferin substrates to luciferin. The luciferin is then measured by luminescence after the addition of a luciferase detection reagent. Pro-luciferin substrate concentration in the assay was equal to its KM for its CYP P450 isozyme. Inhibitors and some substrates limit the production of luciferin and decrease measured luminescence. A compound was classified as a CYP inhibitor if the AC50 (the compound concentration leads to 50% of the activity of an inhibition control) value was 10μM. A compound was considered as a non-inhibitor if AC50 was >57μM. Regarding samples sizes, for CYP1A2 there were 13,256 total compounds with 7,256 non-inhibitors and 6,000 inhibitors, for CYP2C9 there were 12,901 compounds with 8,782 non-inhibitors and 4,119 inhibitors, for CYP2C19 there were 13,445 molecules with 7,532 non-inhibitors and 5,913 inhibitors, for CYP2D6 there were 13,910 compounds 11,139 non-inhibitors and 2,771 inhibitors, and for CYP3A4 there were 13,017 compounds with 7,751 non-inhibitors and 5,266 inhibitors. The chemical name, CAS numbers, CYP test results, and SMILES strings were available in the data.

**Chemical Fingerprints for Toxicity and ADME Predictions.** One approach to computational ADME and toxicity prediction employs chemical substructure analysis of known compounds which have been tested and [67] applies the associative rules between structure and outcome to new compounds whose substructure has been determined. The traditional method for identifying chemical substructure in compounds has been based on the FP2 fingerprint, which yields the presence (absence) of various atoms, bonds, aromaticity and cyclicity, and fine structure of a compound. FP2 fingerprints are in the form of binary 1024-bit vectors which signify presence and absence of the various moieties. It is important to note that while the granularity of FP2 fingerprints

Peer-reviewed version available at Molecules 2019, 24, 645; doi:10.3390/molecules24030645



is high, there is less available information related to copy number of substructure elements, so any exercise is essentially hinged to a binary yes/no dilemma.

Using the toxicity and ADME training data described above, we employed the .NET OB assembly [55] to transform SMILES strings for each training compound into a FP2 1024-bit vector representing chemical substructures. OB yields FP2 fingerprints in the form of 256 4-byte Hex characters were translated to binary bits. Bit values were transformed from 0 to -1, and 1 to 1+ and appended to an analytic file with ADME or toxicity outcomes of the respective training molecule. Toxicity and ADME predictions for the selected DTP ligands were based on trained logistic regression models using 25-100 fingerprints that achieved an ROCAUC>65% for leave-one-out cross validation. Therefore, the predictive results are crude approximations.

Figure 6 illustrates the workflow employed for all ligand preparation, receptor preparation, docking, drug- and lead-like filtering of docked ligands, and computational toxicology and ADME predictions.

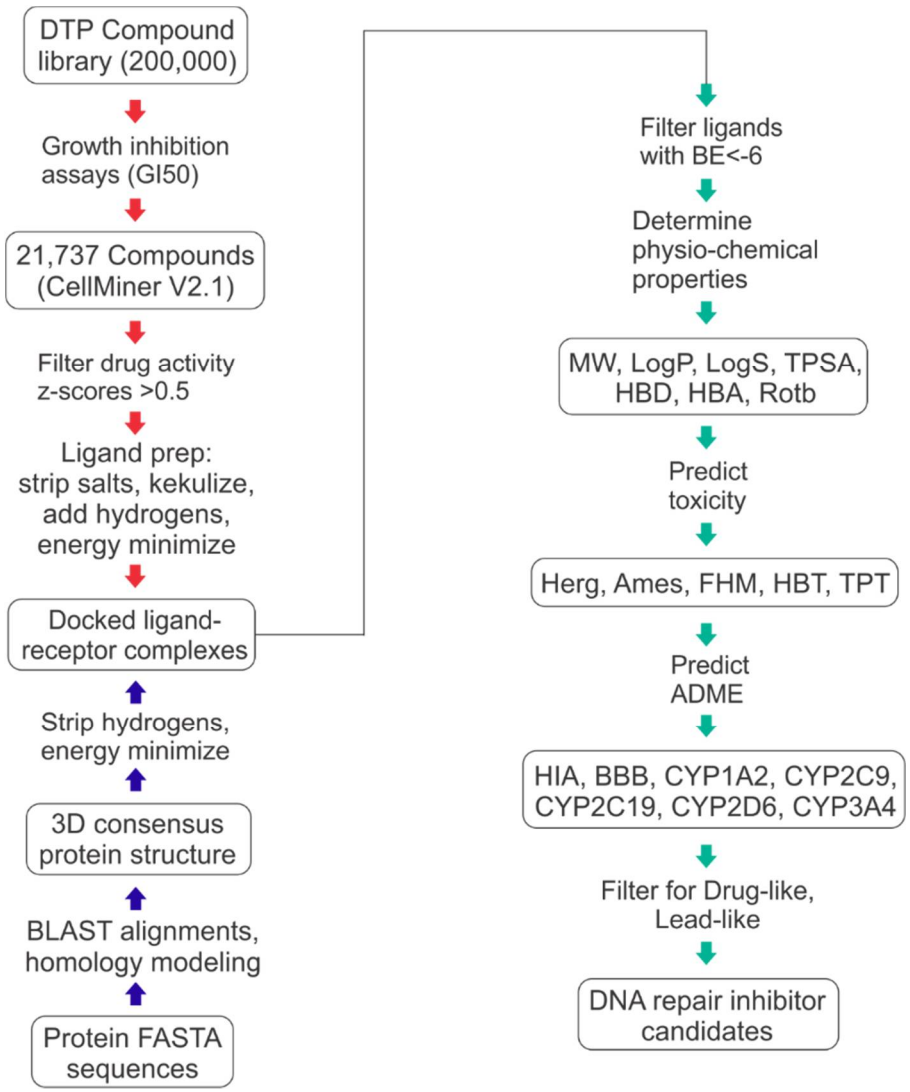



**Figure 6.** Workflow for ligand-receptor docking and drug- and lead-like hit identification.

## 5. Conclusions

We employed computational methods to derive ligand-receptor binding and prediction of toxicological and ADME endpoints for synthetic and natural compounds in the DTP repository and DNA repair proteins which were predictive of OS after adjustment for the PCNA metagene. Results of our computational methods translate to portraits of potentially new repair inhibitors of delayed tumor progression. The utility of our findings can be realized by oncologists and biologists who envision new targets and mechanisms of repair inhibition. We conclude that the results presented can be directly employed for in vitro and in vivo confirmatory experiments to identify new targets of therapy for cancer survival.

**Supplementary Materials:** The following are available online at www.mdpi.com/xxx/s1: Table S1. GBM: List of physio-chemical properties and predicted toxicology and ADME for top 20 ligands. Toxicology and ADME predictions are probabilities in the range [0,1]; Table S2. Melanoma: List of physio-chemical properties and predicted toxicology and ADME for top 20 ligands. Toxicology and ADME predictions are probabilities in the range [0,1]; Table S3. Ovarian: List of physio-chemical properties and predicted toxicology and ADME for top 20 ligands. Toxicology and ADME predictions are probabilities in the range [0,1]; Table S4. Renal papillary: List of physio-chemical properties and predicted toxicology and ADME for top 20 ligands. Toxicology and ADME predictions are probabilities in the range [0,1]; Figure S1. 2D structures of top 20 drug-like ligands for GBM; Figure S2. 2D structures of top 20 drug-like ligands for melanoma; Figure S3. 2D structures of top 20 drug-like ligands for ovarian cancer; Figure S4. 2D structures of top 20 drug-like ligands for renal papillary cancer.

**Funding:** Research was supported by NASA Grant NNX-12AO52A.

**Conflicts of Interest:** The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References


1. Curtin NJ: **DNA repair dysregulation from cancer driver to therapeutic target**. *Nat Rev Cancer* 2012, **12**(12):801-817.
2. Hanahan D, Weinberg RA: **The hallmarks of cancer**. *Cell* 2000, **100**(1):57-70.
3. Cannan WJ, Pederson DS: **Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin**. *J Cell Physiol* 2016, **231**(1):3-14.
4. Forment JV, Kaidi A, Jackson SP: **Chromothripsis and cancer: causes and consequences of chromosome shattering**. *Nat Rev Cancer* 2012, **12**(10):663-670.
5. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, Kyle S, Meuth M, Curtin NJ, Helleday T: **Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase**. *Nature* 2005, **434**(7035):913-917.
6. Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C *et al*: **Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy**. *Nature* 2005, **434**(7035):917-921.
7. Shaheen M, Allen C, Nickoloff JA, Hromas R: **Synthetic lethality: exploiting the addiction of cancer to DNA repair**. *Blood* 2011, **117**(23):6074-6082.
8. Allen C, Ashley AK, Hromas R, Nickoloff JA: **More forks on the road to replication stress recovery**. *J Mol Cell Biol* 2011, **3**(1):4-12.
9. Budzowska M, Kanaar R: **Mechanisms of dealing with DNA damage-induced replication problems**. *Cell Biochem Biophys* 2009, **53**(1):17-31.
10. Nickoloff JA, Jones D, Lee SH, Williamson EA, Hromas R: **Drugging the Cancers Addicted to DNA Repair**. *J Natl Cancer Inst* 2017, **109**(11).
11. Ashworth A, Lord CJ: **Synthetic lethal therapies for cancer: what's next after PARP inhibitors?** *Nat Rev Clin Oncol* 2018, **15**(9):564-576.
12. Dedes KJ, Wilkerson PM, Wetterskog D, Weigelt B, Ashworth A, Reis-Filho JS: **Synthetic lethality of PARP inhibition in cancers lacking BRCA1 and BRCA2 mutations**. *Cell Cycle* 2011, **10**(8):1192-1199.
13. Gavande NS, VanderVere-Carozza PS, Hinshaw HD, Jalal SI, Sears CR, Pawelczak KS, Turchi JJ: **DNA repair targeted therapy: The past or future of cancer treatment?** *Pharmacol Ther* 2016, **160**:65-83.
14. Rehman FL, Lord CJ, Ashworth A: **Synthetic lethal approaches to breast cancer therapy**. *Nat Rev Clin Oncol* 2010, **7**(12):718-724.
15. Aoki D, Chiyoda T: **PARP inhibitors and quality of life in ovarian cancer**. *Lancet Oncol* 2018, **19**(8):1012-1014.
16. Jiang Y, Dai H, Li Y, Yin J, Guo S, Lin SY, McGrail DJ: **PARP inhibitors synergize with gemcitabine by potentiating DNA damage in non-small-cell lung cancer**. *Int J Cancer* 2018.
17. Lyons TG, Robson ME: **Resurrection of PARP Inhibitors in Breast Cancer**. *J Natl Compr Canc Netw* 2018, **16**(9):1150-1156.
18. O'Cearbhaill RE: **Using PARP Inhibitors in Advanced Ovarian Cancer**. *Oncology (Williston Park)* 2018, **32**(7):339-343.
19. Paradiso A, Andreopoulou E, Conte P, Eniu A, Saghatchian M: **PARP Inhibitors in Breast Cancer: Why, How, and When?** *Breast Care (Basel)* 2018, **13**(3):216-219.
20. Robert M, Patsouris A, Frenel JS, Gourmelon C, Augereau P, Campone M: **Emerging PARP inhibitors for treating breast cancer**. *Expert Opin Emerg Drugs* 2018:1-11.

21. Szalat R, Samur MK, Fulciniti M, Lopez M, Nanjappa P, Cleynen A, Wen K, Kumar S, Perini T, Calkins AS *et al*: **Nucleotide excision repair is a potential therapeutic target in multiple myeloma**. *Leukemia* 2018, **32**(1):111-119.
22. Andrews BJ, Turchi JJ: **Development of a high-throughput screen for inhibitors of replication protein A and its role in nucleotide excision repair**. *Mol Cancer Ther* 2004, **3**(4):385-391.
23. Gentile F, Tuszynski JA, Barakat KH: **New design of nucleotide excision repair (NER) inhibitors for combination cancer therapy**. *J Mol Graph Model* 2016, **65**:71-82.
24. Rocha JC, Busatto FF, Guecheva TN, Saffi J: **Role of nucleotide excision repair proteins in response to DNA damage induced by topoisomerase II inhibitors**. *Mutat Res Rev Mutat Res* 2016, **768**:68-77.
25. Zhu G, Myint M, Ang WH, Song L, Lippard SJ: **Monofunctional platinum-DNA adducts are strong inhibitors of transcription and substrates for nucleotide excision repair in live mammalian cells**. *Cancer Res* 2012, **72**(3):790-800.
26. Davidson D, Amrein L, Panasci L, Aloyz R: **Small Molecules, Inhibitors of DNA-PK, Targeting DNA Repair, and Beyond**. *Front Pharmacol* 2013, **4**:5.
27. Mould E, Berry P, Jamieson D, Hill C, Cano C, Tan N, Elliott S, Durkacz B, Newell D, Willmore E: **Identification of dual DNA-PK MDR1 inhibitors for the potentiation of cytotoxic drug activity**. *Biochem Pharmacol* 2014, **88**(1):58-65.
28. Olsen BB, Fritz G, Issinger OG: **Characterization of ATM and DNA-PK wild-type and mutant cell lines upon DSB induction in the presence and absence of CK2 inhibitors**. *Int J Oncol* 2012, **40**(2):592-598.
29. Pospisilova M, Seifrtova M, Rezacova M: **Small molecule inhibitors of DNA-PK for tumor sensitization to anticancer therapy**. *J Physiol Pharmacol* 2017, **68**(3):337-344.
30. Singh SK, Wu W, Zhang L, Klammer H, Wang M, Iliakis G: **Widespread dependence of backup NHEJ on growth state: ramifications for the use of DNA-PK inhibitors**. *Int J Radiat Oncol Biol Phys* 2011, **79**(2):540-548.
31. Ronco C, Martin AR, Demange L, Benhida R: **ATM, ATR, CHK1, CHK2 and WEE1 inhibitors in cancer and cancer stem cells**. *Medchemcomm* 2017, **8**(2):295-319.
32. Sarkaria JN: **Identifying inhibitors of ATM and ATR kinase activities**. *Methods Mol Med* 2003, **85**:49-56.
33. Arnaudeau C, Rozier L, Cazaux C, Defais M, Jenssen D, Helleday T: **RAD51 supports spontaneous non-homologous recombination in mammalian cells, but not the corresponding process induced by topoisomerase inhibitors**. *Nucleic Acids Res* 2001, **29**(3):662-667.
34. Huang F, Motlekar NA, Burgwin CM, Napper AD, Diamond SL, Mazin AV: **Identification of specific inhibitors of human RAD51 recombinase using high-throughput screening**. *ACS Chem Biol* 2011, **6**(6):628-635.
35. Nomme J, Renodon-Corniere A, Asanomi Y, Sakaguchi K, Stasiak AZ, Stasiak A, Norden B, Tran V, Takahashi M: **Design of potent inhibitors of human RAD51 recombinase based on BRC motifs of BRCA2 protein: modeling and experimental validation of a chimera peptide**. *J Med Chem* 2010, **53**(15):5782-5791.
36. Normand A, Riviere E, Renodon-Corniere A: **Identification and characterization of human Rad51 inhibitors by screening of an existing drug library**. *Biochem Pharmacol* 2014, **91**(3):293-300.
37. Reed AM, Fishel ML, Kelley MR: **Small-molecule inhibitors of proteins involved in base excision repair potentiate the anti-tumorigenic effect of existing chemotherapeutics and irradiation**. *Future Oncol* 2009, **5**(5):713-726.

38. Wilson SH, Beard WA, Shock DD, Batra VK, Cavanaugh NA, Prasad R, Hou EW, Liu Y, Asagoshi K, Horton JK *et al*: **Base excision repair and design of small molecule inhibitors of human DNA polymerase beta**. *Cell Mol Life Sci* 2010, **67**(21):3633-3647.
39. Peterson L, Kovyrshina, T.: **DNA Repair Gene Expression Adjusted by the PCNA Metagene Predicts Survival in Multiple Cancers**. In: *BioRxiv*. Obtober 17, 2018 edn; 2018.
40. Venet D, Dumont JE, Detours V: **Most random gene expression signatures are significantly associated with breast cancer outcome**. *PLoS Comput Biol* 2011, **7**(10):e1002240.
41. Driscoll JS: **The preclinical new drug research program of the National Cancer Institute**. *Cancer Treat Rep* 1984, **68**(1):63-76.
42. Rubinstein LV, Shoemaker RH, Paull KD, Simon RM, Tosini S, Skehan P, Scudiero DA, Monks A, Boyd MR: **Comparison of in vitro anticancer-drug-screening data generated with a tetrazolium assay versus a protein assay against a diverse panel of human tumor cell lines**. *J Natl Cancer Inst* 1990, **82**(13):1113-1118.
43. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y: **CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set**. *Cancer Res* 2012, **72**(14):3499-3511.
44. Cellminer hdnngc: 2018, **2.1**.
45. UniProt Consortium T: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2018, **46**(5):2699.
46. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L *et al*: **SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information**. *Nucleic Acids Res* 2014, **42**(Web Server issue):W252-258.
47. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L *et al*: **SWISS-MODEL: homology modelling of protein structures and complexes**. *Nucleic Acids Res* 2018, **46**(W1):W296-W303.
48. Guex N, Peitsch MC, Schwede T: **Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective**. *Electrophoresis* 2009, **30 Suppl 1**:S162-173.
49. Benkert P, Kunzli M, Schwede T: **QMEAN server for protein model quality estimation**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W510-514.
50. Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T: **Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology**. *Sci Rep* 2017, **7**(1):10480.
51. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y *et al*: **BLAST: a more efficient report with usability improvements**. *Nucleic Acids Res* 2013, **41**(Web Server issue):W29-33.
52. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.
53. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A* 1992, **89**(22):10915-10919.
54. Henikoff S, Henikoff JG: **Amino acid substitution matrices**. *Adv Protein Chem* 2000, **54**:73-97.
55. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: An open chemical toolbox**. *J Cheminform* 2011, **3**:33.
56. Gilson MK, Given JA, Bush BL, McCammon JA: **The statistical-thermodynamic basis for computation of binding affinities: a critical review**. *Biophys J* 1997, **72**(3):1047-1069.
57. Halgren TA: **Merck Molecular Force Field. I Basis, Form, Scope, Parameterization, and Performance of MMFF94**. *J Comp Chem* 1996, **17**(5-6):490-515.

58. Trott O, Olson AJ: **AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading**. *J Comput Chem* 2010, **31**(2):455-461.
59. De Winter H: **SILICOS-IT Filter-It**. *http://silicos-itbes3-website-eu-west-1amazonawscom/ (Accession date, 11 July, 2018)* 2018.
60. Muegge I, Heald SL, Brittelli D: **Simple selection criteria for drug-like chemical matter**. *J Med Chem* 2001, **44**(12):1841-1846.
61. Cheng F, Shen, J., Li,W., Lee, P.W., Tang, Y. : **In silico prediction of terrestrial and aquatic toxicities for organic chemicals**. *Chin J Pestic Sci* 2010, **12**:477-488.
62. EPA: **US EPA ECOTOX Database. https://cfpub.epa.gov/ecotox/** (*Accession date, 11 July, 2018)* 2018.
63. Cheng F, Shen, J., Li, W., Yu., Y., Liu, G., Lee, P.W., Tang, Y.: **In silico prediction of Tetrahymena Pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods**. *Chemosphere* 2011, **82**(1):1636-1643.
64. Xue Y, Li H, Ung CY, Yap CW, Chen YZ: **Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods**. *Chem Res Toxicol* 2006, **19**(8):1030-1039.
65. Shen J, Cheng F, Xu Y, Li W, Tang Y: **Estimation of ADME properties with substructure pattern recognition**. *J Chem Inf Model* 2010, **50**(6):1034-1041.
66. NCBI: **NCBI PubChem Database AID-1851. Cytochrome panel assay with activity outcomes. https://pubchem.ncbi.nlm.nih.gov/bioassay/1851/** (*Accession date, 11 July, 2018)* 2018.
67. Zaretzki J, Boehm KM, Swamidass SJ: **Improved Prediction of CYP-Mediated Metabolism with Chemical Fingerprints**. *J Chem Inf Model* 2015, **55**(5):972-982.
68. Sun H: **A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption**. *J Chem Inf Comput Sci* 2004, **44**(2):748-757.
69. Bunin BA: **Increasing the efficiency of small-molecule drug discovery**. *Drug Discov Today* 2003, **8**(18):823-826.