Computational Prediction of Metabolites of Tobacco-Specific Nitrosamines by CYP2A13

Kendall Byler[1], Patrudu Makena[2], G.L. Prasad[2], Jerome Baudry[1*]

[1]The University of Alabama in Huntsville, Department of Biological Sciences. Huntsville, Alabama, USA

[2]Reynolds American Incorporated. Winston-Salem, North Carolina, USA

*Corresponding author. email: jerome.baudry@uah.edu

ABSTRACT

A structure-based computational approach for the prediction of tobacco-specific nitrosamine (TSNA) metabolites by cytochrome P450s has been developed that predicts the known CYP2A13 metabolites of nicotine-derived nitrosamine ketone (NNK), *N*-nitrosonornicotine (NNN), and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL) resulting from hydroxylations and heteroatom oxidations reported in metabolomics literature. This computational approach integrates 1) machine learning models trained on quantum-mechanically-derived molecular surface properties for a set of CYP substrates with known metabolites to predict sites of metabolism across CYP isoforms with 2) the use of ensemble molecular docking to identify which of these predictions are conformationally accessible to the CYP2A13 binding site. This method is generalizable to any CYP isoform for which there is structural information, opening the door to the prediction of P450-based metabolite prediction, as well as prediction and rationalization of metabolomics data.

INTRODUCTION

Smoking-induced diseases such as lung cancer and chronic obstructive pulmonary disease (COPD) are linked to exposure to cigarette toxicants [1]. Cigarette smoke contains more than 7,000 chemical constituents, with some of these designated as Harmful and Potentially Harmful Constituents (HPHCs) by the Food and Drug Administration [2]. Some of these HPHCs are directly carcinogenic, while others are procarcinogenic, requiring metabolic activation through cytochrome P450 (CYP) pathways. Tobacco-specific nitrosamines (TSNAs) are carcinogenic compounds produced by the combustion of tobacco, as well as by tobacco curing and processing, as the oxidation products of nicotine and structurally-related compounds such as nornicotine, anabasine, and anatabine. TSNAs are omnipresent in tobacco products, albeit in different concentrations, depending on the processing and the product. Eight TSNAs are found in tobacco: NNN, NNA, NNK, NNAL, NAT, NAB, iso-NNAL, and iso-NNAC (Figure 1).

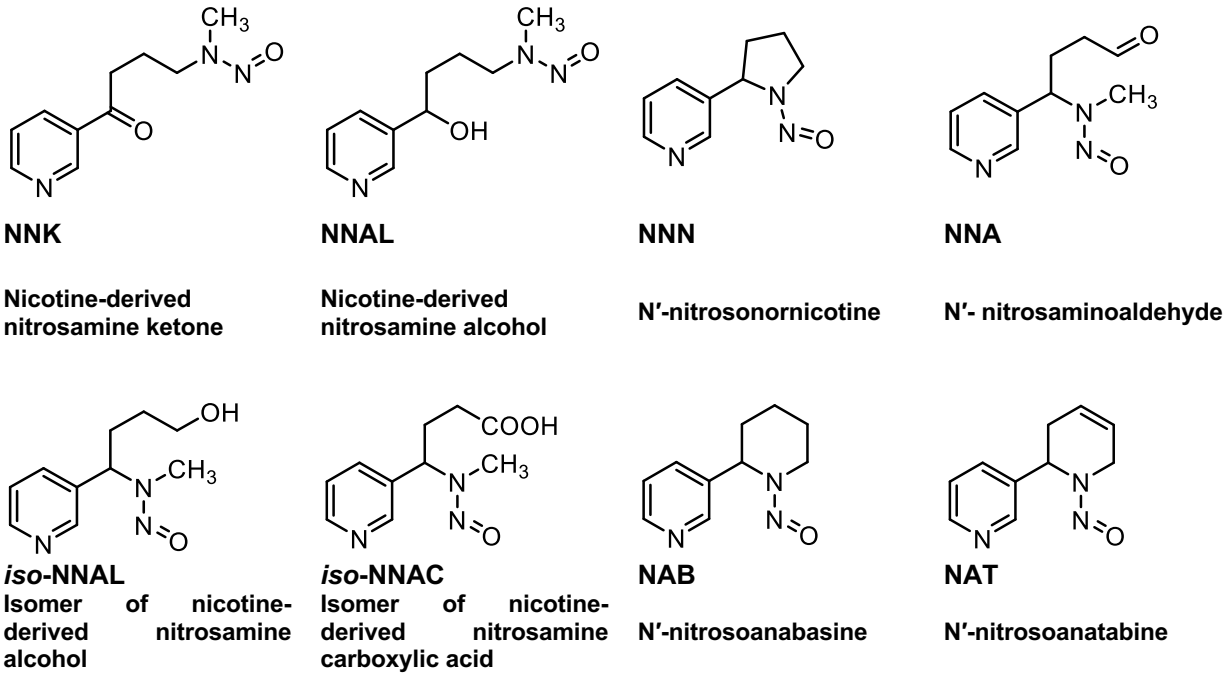
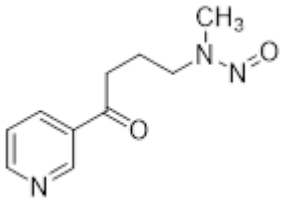


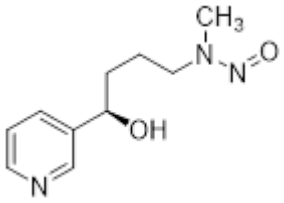
Figure 1 Nitrosamine compounds found in tobacco

Of these, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) N-nitrosonornicotine (NNN), and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL) are the most carcinogenic [3]. The carcinogenic behavior of these nitrosamines occurs as a result of their alkylation of DNA [4]. Damage to DNA accumulates through N- and O-methylation and pyridyloxobutylation of the DNA bases and phosphate groups [5]. Since these tobacco-specific nitrosamines (NNK, *R/S*-NNAL, and NNN) are known to be among the primary carcinogens found in tobacco, the characterization of their CYP metabolites, not only as carcinogenic nitrosamines in themselves, but as biological markers detectable in urine [3], is of particular interest and is the focus of this work on CYP metabolite prediction (Table 1).

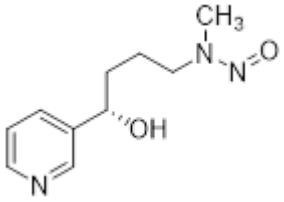
Table 1 TSNAs used in CYP metabolite prediction



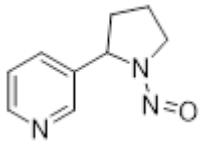
NNK



***R*-NNAL**



***S*-NNAL**



NNN

Cytochrome P450s catalyze several reactions, such as methylations, demethylations, oxidations and hydroxylations, involved in the processing of xenobiotics in preparation for excretion by the kidneys. As hydroxylations and oxidations lead directly to water-soluble metabolites, these are of primary interest in connecting detectable TSNA metabolites to their parent molecules. Several CYPs, including CYP1A1, CYP1B1, CYP2B6, CYP2E1, CYP2J2, CYP2A13, CYP2A6 and CYP3A5, are expressed in human lung [6, 7]. For example, CYP2A6 and CYP2A13 are the most efficient cytochrome P450 enzymes in the metabolic activation of NNK and NNN, with CYP2A13 being 61 – 214 times more efficient than CYP2A6 in the activation of NNK [8, 9]. Some metabolites of HPHCs are well characterized as to their metabolic activation by CYPs. However, other HPHC metabolites and their dose-dependent impact in the different tissues of the body have not yet been fully characterized [10].

Experimental characterization of CYP metabolites of TSNAs is a complex, multidisciplinary process that requires a significant amount of time and effort [1]. To assist in the experimental metabolomics characterization of P450 metabolites, computational modeling approaches have proven to be powerful approaches that have a predictive and rationalizing aspect [11–13]. A computational prediction tool that prioritizes the most likely TSNA CYP metabolites would be extremely helpful in prioritization of the experimental metabolic experiments (*e.g.*, which metabolites are most likely to be produced from a given set of chemicals in a given tissue) and in the analysis of experiments. The present work aims at developing such a predictive tool for the prediction of TSNA metabolites of cytochrome P450s. The present manuscript presents an approach that integrates structural biology (ensemble docking and structure-based properties) together with artificial intelligence (machine learning) to identify the most likely metabolites of TSNAs produced by the human cytochrome P450 isoform CYP2A13, known to efficiently metabolize TSNAs in the lung [14].

Several computational approaches to predicting CYP oxidation sites from chemical structure have already been reported in the literature [15] and are currently available for use through web servers. Two of the more recent applications use quantitative structure-activity relationship (QSAR) modeling of molecular surface properties (CYPScore) [16] and Density Functional Theory (DFT) transition state energies of molecular fragments (SMARTCyp) [17] to predict substrate metobolism. The CYPScore model is trained on substrates with known metabolitic sites produced from metabolism by several CYP isoforms: 3A4, 3A5, 2D6, 2C9, 1A2, 2C19, 2E1, while the original SMARTCyp model was trained using substrates metabolized by 3A4 alone. The predictions made by these approaches, while powerful when dealing with one of these P450s [11], are not applicable to CYP isoforms that have not been fitted to experimental data. And the structural relationship between the CYP-substrate reaction complex and its resulting metabolite are not considered and cannot be recovered. As shown on Figure 2, the approach described here extends beyond these previously-developed computational tools by combining the prediction of isoform-independent CYP metabolism from quantum mechanical surface properties [18] of the substrates (which provides the ability to predict CYP hydroxylations, double bond oxidations, heteroatom oxidations and dealkylations; arrow “A” on Figure 2) with isoform-specific site prediction through the use of ensemble docking (Arrow “B” on Figure 2), which is recognized as a specific and efficient structure-based approach to ligand prediction [19]. This method is thus able to relate a predicted metabolite to a three-dimensional representation of the oxidation reaction that produced it.

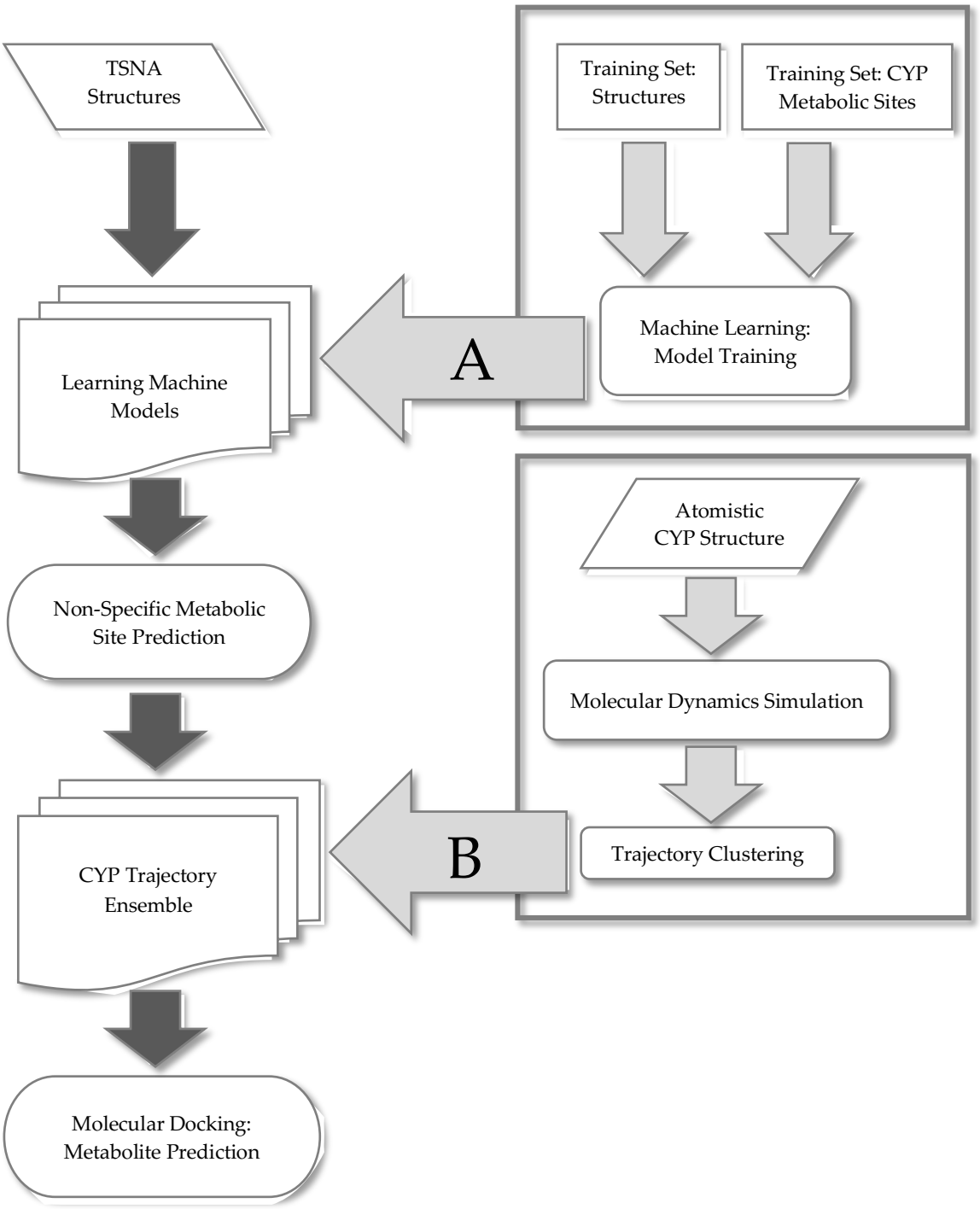


Figure 2 Flowchart of the integration machine learning (A) and ensemble docking (B) in the present computational approach to TSNA metabolite prediction

METHODS

Machine Learning Prediction of CYP Reactivity

A set of 620 known substrates with known CYP metabolism sites, representing 9726 non-hydrogen (heavy) atoms, collected from the literature [16, 17] was used to train machine learning models to identify molecular and structural properties associated with P450 oxidation single-step hydroxylations and N→O oxidations. Cepos InSilico's Parasurf '10 [20] was used to calculate quantum-mechanically-derived molecular surface properties of TSNAs at the AM1 semiempirical level of theory. The local surface properties calculated for each atom in the data set are molecular electrostatic potential, local ionization potential, local electron affinity and local polarizability (Table 2). The Sybyl atom types originally developed by Tripos, Inc., augmented by five new atom types (Supplementary Information) to represent previously non-parametrized atom types, were used to train individual learning machines using the multilayer perceptron, support vector machine and AdaBoost classifiers in the SciKit-Learn package [21]. Learning machines of each type were trained using the local properties, and including the AM1 Mulliken partial charge, for each atom type. For each atom type, the learning machine with the best classification statistics was used to predict activity for that atom type. Relevant machine learning classification rates are given in Supplementary Information.

Table 2 Local molecular surface properties used in addition to Mulliken atomic partial charge to predict active sites

Local Property	Description
SA	Solvent Accessible Surface Area
MEP max	Molecular Electrostatic Potential maximum
MEP min	Molecular Electrostatic Potential minimum
IE_l max	Local Ionization Potential maximum
IE_l min	Local Ionization Potential minimum
EA_l max	Local Electron Affinity maximum
EA_l min	Local Electron Affinity minimum
POL_{mean}	Mean Polarizability
Field(N) max	Electrostatic Field maximum
Field(N) min	Electrostatic Field minimum

Structure-Based Prediction of CYP Reactivity

The TSNA compounds used in the present work were nicotine-derived nitrosamine ketone (NNK), *N*-nitrosonornicotine (NNN), and the *R* and *S* enantiomers of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL), as shown in Table 1. The molecular docking calculations were performed using the docking facility in MOE2019.01 [22] with the London ΔG scoring function for initial placements and the GBVI/WSA ΔG scoring function during the refinement steps. The force field used for refinement was Amber 10:EHT with Generalized Born approximations to Poisson-Boltzmann electrostatics. The CYP2A13 X-ray crystal structure used in this work was the human lung CYP2A13 co-crystallized with nicotine (PDB ID: 4EJG). Protonation states were assigned at pH 7 using the Protonate3D facility in MOE.

The protonated structure of CYP2A13 was used as the starting structure for the 100 ns molecular dynamics simulation using NAMD [23] with Amber10 force field parameters, along with heme parameters published by Shahrokh, *et al.* [24] In preparing the simulation, six TIP3 water molecules were automatically placed into the binding pocket volume previously occupied by nicotine in 4EJG by automated solvation in MOE and their orientations in the pocket optimized. The explicitly-solvated (TIP3) periodic system was generated with 10 Å padding on each side of the system and equilibrated for 100 ps, followed by a production run at 310 K and 1 bar using a Langevin barostat and thermostat. The resulting CYP conformations were used in ensemble docking of the TSNAs in the active site. Although ensemble docking provides an avenue for treating the target as a flexible entity, it typically requires the use of supercomputing facilities to perform the long MD simulations and extensive numbers of dockings. For instance, in the case of the 2A13 system modeled here, 100 ns of simulation time generates approximately 50,000 frames – each of which could potentially be used as targets to which the TSNA set would need to be docked, leading to roughly half a million poses per TSNA that would need to be evaluated. To run the calculations on a non-supercomputer machine in a reasonable amount of time required some trade-off between the dynamic treatment of the binding site along with a significant reduction in the number of docking calculations. As suggested in [11, 19, 25–27], the MD trajectory were clustered by the RMSD of binding site residues so that a diverse set of binding site conformations is selected for docking. This clustering was performed using the binding site residues Phe 107, Ala 117, Phe 118, Phe 209, Phe 300, Ala 301, Thr 305, Leu 366 and Leu 370 with Chimera, which also identified representative frames for each cluster (the trajectory frame closest to the average of each cluster), and sorted by the occupancy of the cluster (Figure 3). As in previous ensemble

docking approaches, a cutoff of 10 members per cluster was chosen, resulting in a set of ten representative frames of this MD trajectory.

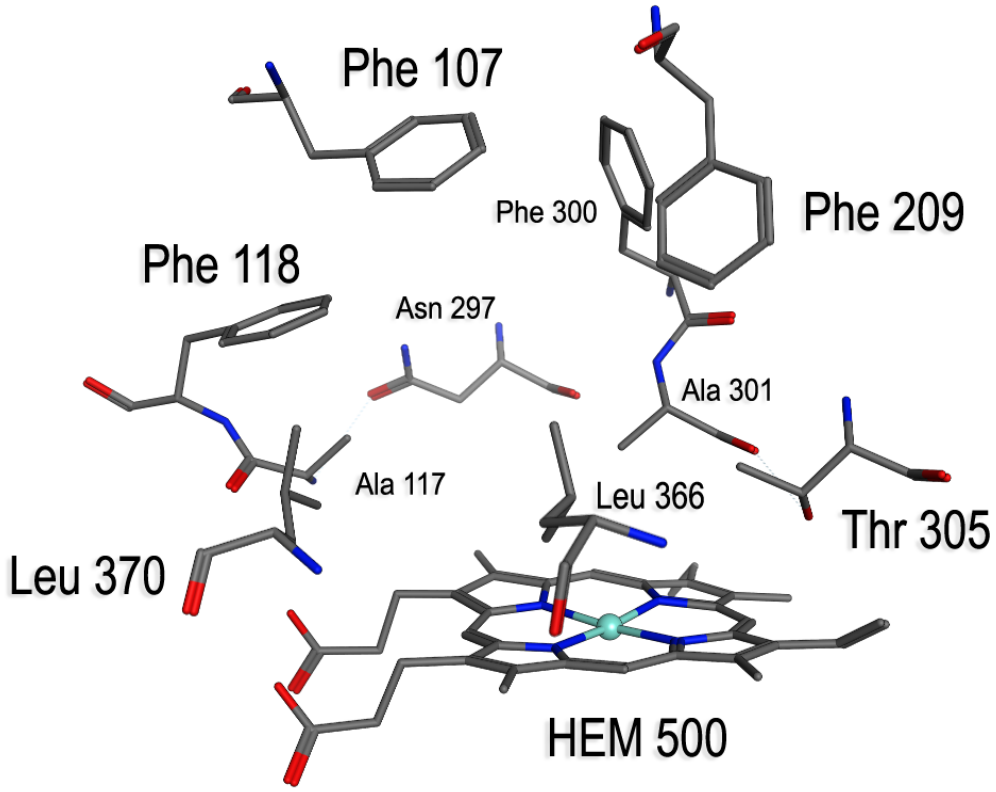


Figure 3 2A13 binding site residues used for dynamics trajectory clustering

A proximity cutoff for a substrate's heavy atoms to the virtual oxo-heme oxygen was set at 4.0 Å, which was found necessary to recover all TSNA active sites in the subsequent dockings, in agreement with reactive distances (3.0 Å – 5.2 Å) for CYPs used in previous published computational CYP-substrate modeling work [11, 28]. Over the course of the molecular dynamics simulation, an *average* of 6 waters remained within 10 Å of the heme cofactor.

Each TSNA was docked to the binding sites of the 10 representative frames of the CYP2A13 MD trajectory ensemble, retaining the top 50 poses in each case. While the TSNA structures were docked to the reduced species (*i.e.* having no oxygen bound to the heme iron of the CYP structure), prior *ab initio* modeling [24] of the Fe^{IV}=O complex suggests that the activated oxygen is 1.694 Å away from the iron. Accordingly, prior to docking, the coordinates of a virtual oxygen

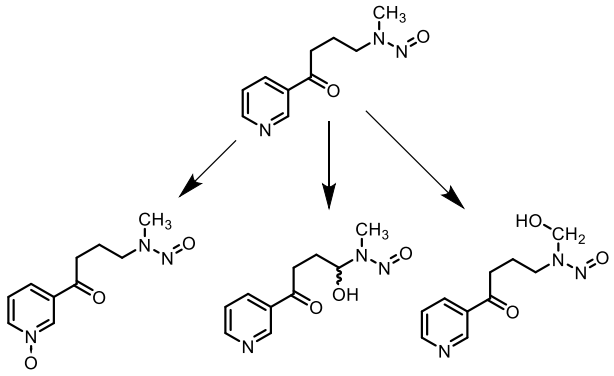
attached to the heme iron were calculated for each frame to be in 1.694 Å away from the heme's iron. In order for a heavy atom to be considered as the site of oxidation by the docking calculations, it had to fulfill both of the following criteria: 1) the atom is the closest heavy atom within a cutoff distance of 4.0 Å to the virtual oxygen's atomic location in a given docked pose, and 2) that atom has also been identified as an active site by the learning machines. Poses that did not meet this criterion were not considered for analysis.

RESULTS AND DISCUSSION

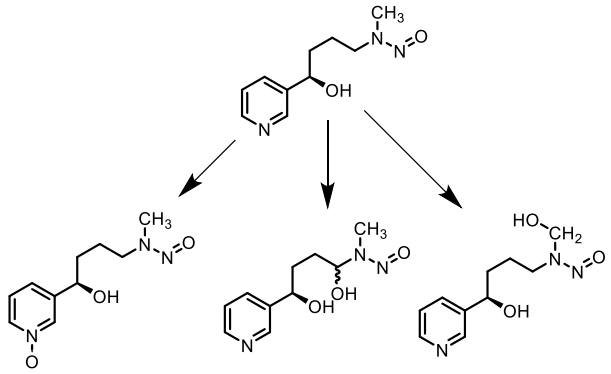
The prediction tool presented here focuses specifically on the prediction of TSNA metabolites that arise from single-step hydroxylations and N→O oxidations by CYPs, as shown in Table 3. The depicted metabolite structures represent the first oxidations associated with known CYP oxidative pathways that lead to other products detected in metabolomics studies.

Table 3 Known metabolites of the four TSNAs used in this study

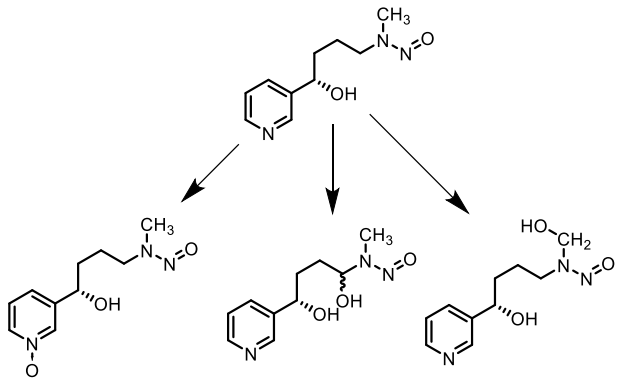
NNK



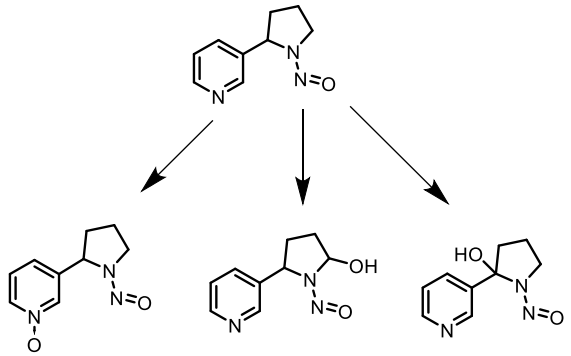
R-NNAL



S-NNAL



NNN



Machine learning predictions

Table 4 shows the predicted hydroxylation and nitrogen oxidation sites for all four parent TSNA molecules evaluated here using the machine learning approach. None of the 4 TSNAs were included in the machine learning training set. In each case, a fourth site of CYP interaction was predicted involves the formation of a coordinate covalent bond between the TSNA nitrosamine nitrogen and the iron in the CYP heme (indicated in the crystal structure of the NNK-2A13 coordination complex in [29], PDB ID: 4EJH), rather than a pathway leading to a known metabolite. The machine learning models described in Methods correctly, and uniquely, identified each of the known oxidation sites for each TSNA shown in Table 3, representing the known α-hydroxylation products of each TSNA.

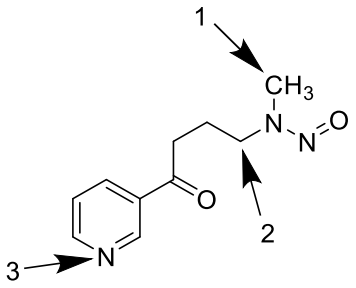
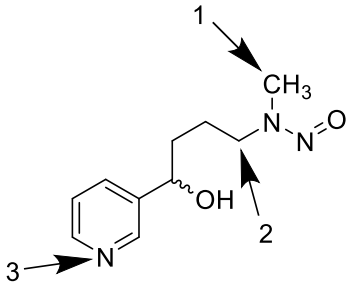
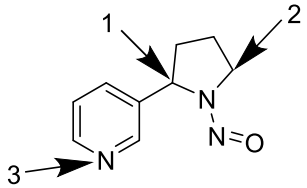
TSNA	Sites of CYP Metabolism
NNK	
***R*-NNAL** ***S*-NNAL**	
NNN	

Figure 4 Tobacco-Specific Nitrosamine metabolic sites predicted by the machine learning computational model

Docking-based predictions

After clustering the 2A13 MD trajectory, each of the ten frames was used to dock the TSNA set using the parameters described in Methods, resulting in a combined list of 500 poses for each TSNA. After ranking by docking score, all poses with an atomic site selected as a site of CYP metabolism by the machine learning models within the 4.0 Å cutoff from the virtual heme oxygen in each ensemble frame were compiled for each TSNA, as described in Methods, resulting in much shorter pose lists. The results obtained from ensemble docking are summarized below in Table 5. Of the 298 TSNA docking poses identified by docking, 37 were NNK poses, 53 were *R*-NNAL poses, 53 were *S*-NNAL poses, and 155 were NNN poses. And for or all TSNAs but NNK, the most frequently selected active sites from TSNA docking to the 2A13 ensemble were the aromatic pyridine nitrogens (pathways leading to N→O oxidation metabolites), representing an average of 44% of all recovered poses. This latter finding is not in accord with published literature on 2A13 metabolites of NNN and NNK [30]. Indeed, the most energetically favorable poses from the molecular dockings are those associated with reaction complexes leading to N→O oxidations. We attribute this anomaly to the fact that heme oxygens were not modeled in the dynamics simulation and are thus not represented in the docking ensemble. The Coulombic repulsions due to the negative partial atomic charges calculated for both the TSNA pyridine nitrogens and the heme oxygens would substantially lower the predicted binding free energies in the docking screen, and relegate these poses to the bottom of the ranked pose list – and most likely beyond the threshold set for our selection criteria.

Table 5 Number of docking poses needed to recover all sites identified as active by machine learning models

TSNA	Minimum number of poses recovering all active sites	Number of poses that met selection criteria
NNK	19	37
***R*-NNAL**	23	53
***S*-NNAL**	25	53
NNN	27	155

In future developments of this approach, the pose list, ranked by docking score for each active pose, could potentially be used as an indicator of the relative distribution of metabolite structures

produced by a given CYP. This last step would require additional modeling of available metabolomics data for known substrates in order to relate the relative distribution of each metabolite to the number of poses indicated for that metabolite, and is outside the scope of the present paper. A compelling reason for not including metabolite concentration prediction in the current model stems from the challenge presented by often conflicting information found in literature sources regarding metabolite detection. For instance, in metabolomics studies that report no detection of N-oxides from CYP metabolism, they may have been mis-assigned in the mass spectroscopic analysis: “The identification of N-oxides represents a challenge because both hydroxylation and N-oxidation result in an increase in molecular weight by 16. The molecular ions of these metabolites are indistinguishable by mass spectrometry” [31]. Particularly in the cases of N-oxide metabolites, we expect there to be discrepancies among studies reporting the absence or presence of detectable N-oxides and that this contributes to a crisis of source material because the CYP N-oxide metabolites of TSNAs are reported in the literature: “Similarly, NNK-N-oxide is the major metabolite formed in the isolated and perfused rat lung, but not in the isolated and perfused rat liver. NNK-N-oxide is a major metabolite formed by patas monkey lung microsomes, but is not formed in by liver microsomes. In contrast to both rodents and the patas monkey, pyridine-N-oxidation of NNK and NNAL are observed in human liver microsomes, but not in human lung microsomes.” [32].

Table 6 Calculated binding site volumes and number of correctly-identified TSNA poses in the 2A13 ensemble of 10 structures used in ensemble docking.

	Binding Site Volume (Å^3)	**Number of TSNA Poses fulfilling the Methods criteria**
1)	20.992	49
2)	27.648	28
3)	24.576	14
4)	19.968	31
5)	39.424	28
6)	30.720	12
7)	5.120	19

8)	11.264	19
9)	37.376	25
10)	32.768	22

As shown in Table 6, the calculated binding site volumes for the individual frames of the MD ensemble does not correlate well with the number of docking poses selected, suggesting that the shape and chemical properties of the binding site, as determined by the relative conformations of the binding site residues, is largely responsible for the selectivity in identifying metabolically active poses, rather than only a steric effect.

CONCLUSIONS

The computational model presented here identifies known hydroxylation and N→O oxidation CYP2A13 metabolites in a structure-dependent manner for the four TSNA substrates evaluated: nicotine-derived nitrosamine ketone (NNK), *N*-nitrosonornicotine (NNN), and the *R* and *S* enantiomers of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL). The machine learning models in this method are also trained to identify sites of CYP metabolism associated with aliphatic, aromatic and benzylic hydroxylations, double bond oxidations, other heteroatom oxidations and heteroatom dealkylations. The present approach may be used to predict the metabolites of other substrates in the future. Since CYPs are known to metabolize a wide variety of substrate structures, their binding sites may be highly flexible. Ensemble docking selects the machine-learning-predicted metabolic sites that are within a reactive distance of the heme moiety in the CYP binding pocket. This work opens the door to the systematic prediction of metabolites for a variety of substrates by any given CYP isoform and, potentially, of the relative abundances of specific metabolites by fitting to experimental data.

REFERENCES

1. Hecht SS, Hoffmann D. Tobacco-specific nitrosamines, an important group of carcinogens in tobacco and tobacco smoke. *Carcinogenesis*. 1988, *9*, 875–84.

2. Harmful and Potentially Harmful Constituents in Tobacco Products and Tobacco Smoke: Established List. The U.S. Food & Drug Administration. 2012. https://www.fda.gov/tobacco-products/rules-regulations-and-guidance/harmful-and-potentially-harmful-constituents-tobacco-products-and-tobacco-smoke-established-list.

3. Hecht SS. Biochemistry, biology, and carcinogenicity of tobacco-specific N- nitrosamines. *Chem Res Toxicol*. 1998, *11*, 559–603.

4. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2010.

5. Hecht SS. DNA adduct formation from tobacco-specific N-nitrosamines. *Mutat Res - Fundam Mol Mech Mutagen*. 1999, *424*, 127–42.

6. Lu W, Banerjee B, Molland KL, Seleem MN, Ghafoor A, Hamed MI, et al. Synthesis of 3-(3-aryl-pyrrolidin-1-yl)-5-aryl-1,2,4-triazines that have antibacterial activity and also inhibit inorganic pyrophosphatase. *Bioorganic Med Chem*. 2014, *22*, 406–18.

7. Zhang X, Agostino J, Wu H, Zhang Q-Y, von Weymarn L, Murphy SE, et al. CYP2A13: Variable Expression and Role in Human Lung Microsomal Metabolic Activation of the Tobacco-Specific Carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. *J Pharmacol Exp Ther*. 2007, *323*, 570–8. doi:10.1124/jpet.107.127068.

8. He X-Y, Shen J, Ding X, Lu AYH, Hong J-Y. Identification of Critical Amino Acid Residues of Human CYP2A13 for the Metabolic Activation of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a Tobacco-Specific Carcinogen. *Drug Metab Dispos*. 2004, *32*, 1516–21. doi:10.1124/dmd.104.001370.

9. Su T, Bao Z, Zhang Q-Y, Smith TJ, Hong J-Y, Ding X. Human Cytochrome P450 CYP2A13: Predominant Expression in the Respiratory Tract and Its High Efficiency Metabolic Activation of a Tobacco-specific Carcinogen, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. *Cancer Res*. 2000, *60*, 5074–9. http://cancerres.aacrjournals.org/content/60/18/5074.abstract.

10. Cunningham A, Sommarström J, Sisodiya AS, Errington G, Prasad K. Longitudinal study of long-term smoking behaviour by biomarker-supported determination of exposure to smoke. *BMC Public Health*. 2014, *14*, 348. doi:10.1186/1471-2458-14-348.

11. Harris JB, Eldridge ML, Sayler G, Menn FM, Layton AC, Baudry J. A computational approach predicting CYP450 metabolism and estrogenic activity of an endocrine disrupting compound (PCB-30). *Environ Toxicol Chem*. 2014, *33*, 1615–23.

12. Baudry J, Rupasinghe S, Schuler MA. Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s. *Protein Eng Des Sel*. 2006, *19*, 345–53. doi:10.1093/protein/gzl012.

13. Li X, Baudry J, Berenbaum MR, Schuler MA. Structural and functional divergence of insect CYP6B proteins: From specialist to generalist cytochrome P450. *Proc Natl Acad Sci U S A*. 2004, *101*, 2939–44. doi:10.1073/pnas.0308691101.

14. Su T, Bao Z, Zhang Q-Y, Smith TJ, Hong J-Y, Ding X. Human Cytochrome P450 CYP2A13: Predominant Expression in the Respiratory Tract and Its High Efficiency Metabolic Activation of a Tobacco-specific Carcinogen, 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone. *Cancer Res*. 2000, *60*, 5074 LP – 5079.

15. Crivori P, Poggesi I. Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *Eur J Med Chem*. 2006, *41*, 795–808. doi:10.1016/J.EJMECH.2006.03.003.

16. Hennemann M, Friedl A, Lobell M, Keldenich J, Hillisch A, Clark T, et al. CypScore: Quantitative Prediction of Reactivity toward Cytochromes P450 Based on Semiempirical Molecular Orbital Theory. *ChemMedChem*. 2009, *4*, 657–69. doi:10.1002/cmdc.200800384.

17. Rydberg P, Gloriam DE, Zaretzki J, Breneman C, Olsen L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med Chem Lett*. 2010, *1*, 96–100. doi:10.1021/ml100016x.

18. Byler K. 3D-QSAR and Physical Property Modeling Using Quantum-Mechanically- Derived Molecular Surface Properties. Friedrich-Alexander-Universität Erlangen-Nürnberg; 2007.

19. Ellingson SR, Miao Y, Baudry J, Smith JC. Multi-Conformer Ensemble Docking to Difficult Protein Targets. *J Phys Chem B*. 2015, *119*, 1026–34. doi:10.1021/jp506511p.

20. El Kerdawy A, Güssregen S, Matter H, Hennemann M, Clark T. Quantum Mechanics-Based

Properties for 3D-QSAR. *J Chem Inf Model*. 2013, *53*, 1486–502. doi:10.1021/ci400181b.

21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011, *12*, 2825–30.

22. Molecular Operating Environment (MOE) 2019, Chemical Computing Group. 2019.

23. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. *J Comput Chem*. 2005, *26*, 1781–802. doi:10.1002/jcc.20289.

24. Shahrokh K, Orendt A, Yost GS, Cheatham TE. Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J Comput Chem*. 2012, *33*, 119–33.

25. Evangelista Falcon W, Ellingson SR, Smith JC, Baudry J. Ensemble Docking in Drug Discovery: How Many Protein Configurations from Molecular Dynamics Simulations are Needed To Reproduce Known Ligand Binding? *J Phys Chem B*. 2019, *123*, 5189–95. doi:10.1021/acs.jpcb.8b11491.

26. Ellingson SR, Smith JC, Baudry J. VinaMPI: Facilitating multiple receptor high-throughput virtual docking on high-performance computers. *J Comput Chem*. 2013, *34*, 2212–21. doi:10.1002/jcc.23367.

27. Amaro RE, Baudry J, Chodera J, Demir Ö, McCammon JA, Miao Y, et al. Ensemble Docking in Drug Discovery. *Biophys J*. 2018, *114*, 2271–8. doi:https://doi.org/10.1016/j.bpj.2018.02.038.

28. Baudry J, Li W, Pan L, Berenbaum MR, Schuler MA. Molecular docking of substrates and inhibitors in the catalytic site of CYP6B1, an insect cytochrome P450 monooxygenase. *Protein Eng Des Sel*. 2003, *16*, 577–87. doi:10.1093/protein/gzg075.

29. DeVore NM, Scott EE. Nicotine and 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone Binding and Access Channel in Human Cytochrome P450 2A6 and 2A13 Enzymes. *J Biol Chem*. 2012, *287*, 26576–85. doi:10.1074/JBC.M112.372813.

30. Liu X, Zhang J, Wang L, Yang B, Zhang C, Liu W, et al. In vitro metabolism of N′-Nitrosonornicotine catalyzed by cytochrome P450 2A13 and its inhibition by nicotine, N′-Nitrosoanatabine and N′-Nitrosoanabasine. *Chem Biol Interact*. 2016, *260*, 263–9.

31. Ramanathan R, Su AD, Alvarez N, Blumenkrantz N, Chowdhury SK, Alton K, et al. Liquid

chromatography/mass spectrometry methods for distinguishing N- oxides from hydroxylated compounds. *Anal Chem*. 2000, *72*, 1352–9.

32. Hecht SS, Tricker AR. Chapter 11 - Nitrosamines derived from nicotine and other tobacco alkaloids. In: Gorrod JW, Jacob PBT-AD of N and RC and their M, editors. Amsterdam: Elsevier Science; 1999. p. 421–88. doi:https://doi.org/10.1016/B978-044450095-3/50012-7.

SUPPLEMENTARY INFORMATION

Table S.1 The set of extended Sybyl atom types used in machine learning training and prediction

C.3	sp^3 hybridized carbons	**N.3**	sp^3 hybridized nitrogens
C.2	nonaromatic sp^2 hybridized carbons	**N.ar**	aromatic nitrogens
C.1	*sp* hybridized carbons	**N.pl3**	trigonal planar nitrogens
C.ar	aromatic sp^2 hybridized carbons	**S.3**	sp^3 hybridized sulfurs
C.am	amide carbons	**S.2**	sp^2 hybridized sulfurs
C.bz	benzylic carbons	**P.3**	sp^3 hybridized phosphorus
C.oa	carbons alpha to oxygen		
C.na	carbons alpha to nitrogen		

Table S.2 Machine learning models used in substrate active site prediction

Atom Type	Cases	Metabolic Sites	Model Type	Avg. Classification Rate (%)*
C.3	1756	178	MLP	81.7
C.2	814	35	MLP	93.0
C.1	22	2	SVC	95.0
C.ar	5374	217	MLP	94.7
C.am	247	3	SVC	99.6
C.bz	412	96	ADA	68.9
C.na	745	281	SVC	70.2
C.oa	670	141	MLP	64.0
N.3	279	13	MLP	90.6
N.ar	391	13	MLP	93.6
N.pl3	184	12	ADA	92.2
S.3	60	44	MLP	74.4
S.2	27	9	SVC	61.7
P.3	12	4	ADA	100.0**

*Ten-fold cross validation using randomly selected 75% training, 25% testing.
**Five-fold cross validation using randomly selected 75% training, 25% testing.
SVC: Support Vector Machine

MLP: Multilayer Perceptron
ADA: AdaBoost Classifier

Table S.3 Trajectory ensemble from the molecular dynamics simulation of 2A13

Cluster	Frame	Members	Cluster	Frame	Members
1)	11005	27	**6)**	11887	18
2)	7897	26	**7)**	1093	17
3)	4831	21	**8)**	3319	15
4)	2689	19	**9)**	5713	12
5)	10165	19	**10)**	8569	10

Table S.4 Number of NNK active sites recovered in the 2A13 ensemble

	Site		
Cluster	**1**	**2**	**3**
1)	0	9	1
2)	1	0	4
3)	0	2	2
4)	0	0	2
5)	0	8	0
6)	0	0	2
7)	0	2	0
8)	0	0	0
9)	0	0	3
10)	0	0	0

Table S.5 Number of *R*-NNAL active sites recovered in the 2A13 ensemble

	Site		
Cluster	**1**	**2**	**3**
1)	0	7	6
2)	1	2	9
3)	0	1	1
4)	1	1	3
5)	1	4	1
6)	0	3	0
7)	0	0	0
8)	0	1	0
9)	2	0	3
10)	1	0	2

Table S.6 Number of *S*-NNAL active sites recovered in the 2A13 ensemble

	Site		
Cluster	**1**	**2**	**3**
1)	2	9	2
2)	1	0	7

3)	0	0	1
4)	0	1	2
5)	0	6	2
6)	0	0	1
7)	0	5	1
8)	0	1	1
9)	2	0	0
10)	1	0	5

Table S.7 Number of NNN active sites recovered in the 2A13 ensemble

	Site		
Cluster	**1**	**2**	**3**
1)	2	4	7
2)	0	0	3
3)	4	1	2
4)	7	7	7
5)	1	1	4
6)	0	2	4
7)	3	4	4
8)	12	2	2
9)	1	5	9
10)	1	7	5

Table S.8 Number of active TSNA poses identified in the 2A13 ensemble

	NNK	**R-NNAL**	**S-NNAL**	**NNN**
1)	10	13	13	13

2)	5	12	8	3
3)	4	2	1	7
4)	2	5	3	21
5)	8	6	8	6
6)	2	3	1	6
7)	2	0	6	11
8)	0	1	2	16
9)	3	5	2	15
10)	0	3	6	13

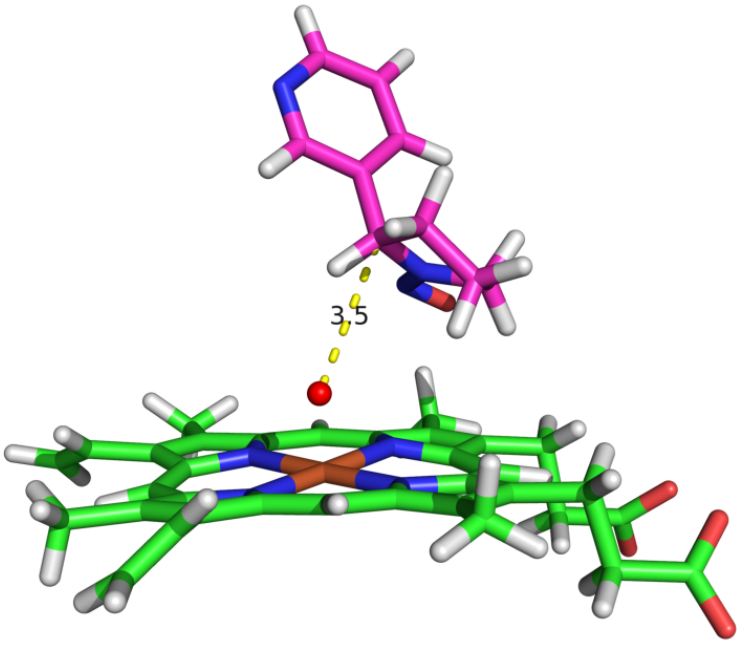


Figure 4 Docking pose of NNN active site 1 with a 4.0 Å proximity to the active heme oxygen of 2A13

ADDENDUM

Since the relative distribution of metabolites resulting from the CYP metabolism for some substrates is available in the literature, it may be possible to extrapolate metabolite distribution from the number of predicted metabolite structures that appear in the ensemble pose lists of this method. As shown in Figure 1, we observed that the binding site cavities of the 2A13 ensemble yielded many more poses associated with NNN metabolites than with the other three TSNAs. Of the 298 TSNA docking poses within 4.0 Å of the virtual heme oxygen in the 2A13 ensemble, 37 were NNK poses, 53 were *R*-NNAL poses, 53 were *S*-NNAL poses, and 155 were NNN poses. And for or all TSNAs but NNK, the most frequently selected active sites from TSNA docking to the 2A13 ensemble were the aromatic pyridine nitrogens (pathways leading to N→O oxidation metabolites), representing an average of 44% of poses (Figure 2).

Figure 1 Distribution of active docking poses for each TSNA in the 2A13 ensemble

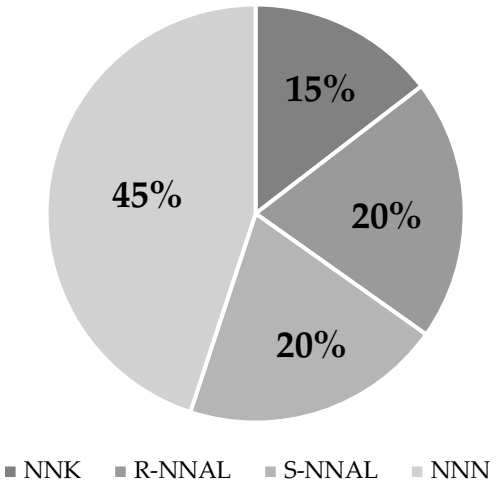
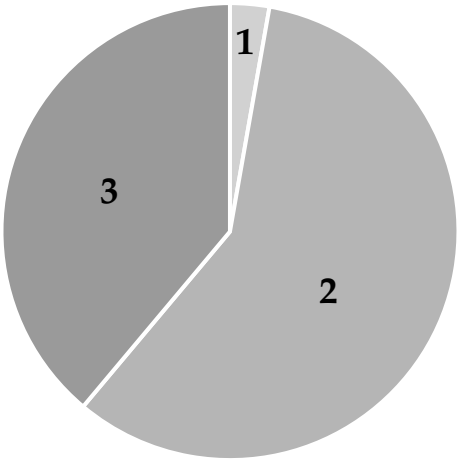
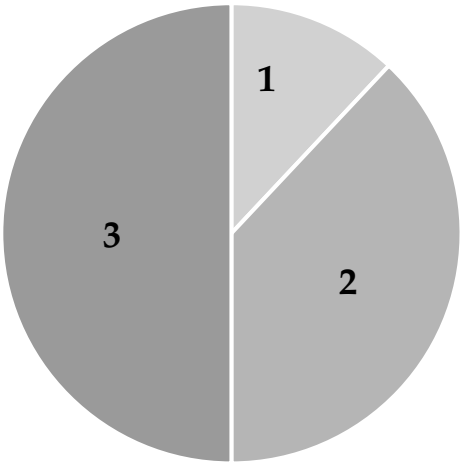


Figure 2 Proportion of active sites predicted for each TSNA in the 2A13 ensemble



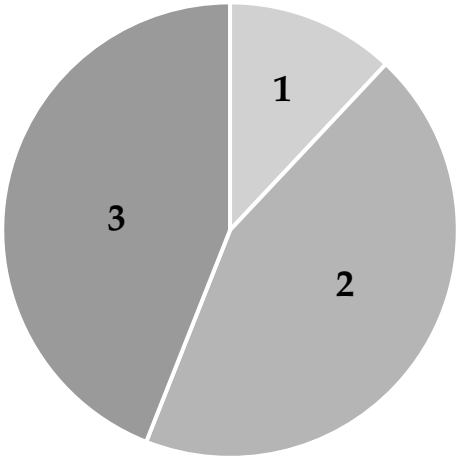
NNK

1	3%
2	58%
3	39%



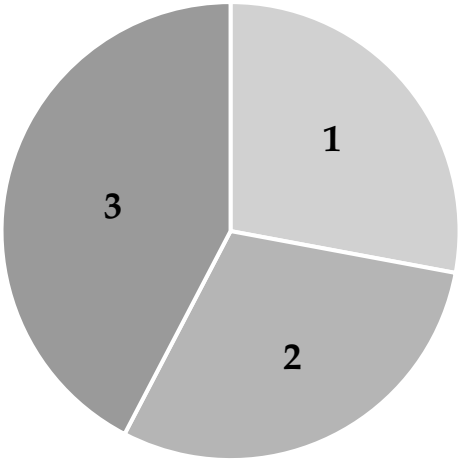
***R*-NNAL**

1	12%
2	38%
3	50%



***S*-NNAL**

1	12%
2	44%
3	44%



NNN

1	28%
2	30%
3	42%