CHAPTER THREE

A structure-guided computational screening approach for predicting plant enzyme–metabolite interactions

Cynthia K. Holland* **and Hisham Tadfie**
Department of Biology, Williams College, Williamstown, MA, United States
*Corresponding author: e-mail address: ckh2@williams.edu

Contents

Abstract

Plants are molecular factories that have spent millions of years evolving the enzymes needed to synthesize diverse primary and specialized metabolites. Despite the wealth of metabolites that plants produce, many of the enzymes responsible for generating these molecules have yet to be identified. For enzymes with known substrates, the extent of substrate promiscuity and small-molecule regulation remains unexplored. Many computational methods for identifying metabolic enzymes focus on gene-based approaches that rely on transcriptomics, metabolomics, and comparative genomics. With new AI-based tools for accurate protein structure prediction, protein-based

Methods in Enzymology, Volume 676
ISSN 0076-6879
https://doi.org/10.1016/bs.mie.2022.07.025

strategies that screen a library of small molecules against a high-quality protein model can facilitate the identification of substrates, products, or inhibitors. Virtual screening has been used for structure-based drug design in the pharmaceutical industry for decades and easily translates to investigating plant metabolic enzymes. Here, we present a method for rapid, user-friendly, and open-source virtual screening using the *Arabidopsis thaliana* UGT74F2 with a curated library of specialized metabolites and herbicides and AutoDock Vina as an example. This method may be applied broadly to metabolic enzymes, and compound libraries can be easily adapted. Compounds are ranked based on their relative binding affinities and the resulting binding modes are evaluated using a molecular visualization program, like PyMOL. Because this is a computational approach, results from the virtual screen will need to be validated using in vitro or in vivo activity, binding, or inhibition assays. Virtual screening may aid in identifying substrates for enzymes of unknown function, revisiting substrate selectivity, or identifying natural or synthetic inhibitors.

1. Introduction

Plants produce as many as a million compounds with diverse roles in growth and development, defense against pathogens and herbivores, protection from abiotic stresses, pollinator attraction, and reproduction (Fang, Fernie, & Luo, 2019). Over millions of years, plants have evolved enzymes that synthesize these specialized metabolites, many of which are the result of gene duplication followed by neofunctionalization of genes from primary metabolism (Maeda & Fernie, 2021). Because many of these compounds are useful to humans for their roles in medicine, nutrition, or crop improvement, researchers have spent the past few decades searching for enzymes that produce these specialized metabolites. As the cost of whole-genome sequencing and transcriptome sequencing has decreased, identification and functional characterization of plant metabolic enzymes has advanced rapidly in recent years. Computational identification of candidate genes coupled to functional assays has led to the identification of entire biosynthetic pathways. While many important specialized metabolic pathways have been elucidated, including the pathways for pharmaceuticals such as morphine, etoposide, and vinblastine, the enzymes involved in synthesizing many pharmaceutically and agriculturally relevant metabolites remain to be identified (Caputi et al., 2018; Schultz, Kim, Lau, & Sattely, 2019; Singh, Menéndez-Perdomo, & Facchini, 2019). To build on existing computational approaches that have been successful in identifying candidate

genes, a protein-based approach to predict the function of enzymes is now possible due to recent advances in AI-based structural protein modeling (Jumper et al., 2021).

1.1 Identifying candidate genes using computational and systems biology approaches

In the two decades since the *Arabidopsis thaliana* genome was completed, over a thousand plant genomes have been sequenced, due in great part to advances in next-generation sequencing such as short- and long-read sequencing and the decreasing costs of whole-genome sequencing (Sun, Shang, Zhu, Fan, & Guo, 2022). Despite the availability of this data, only a fraction of plant genes have been experimentally characterized and many enzymes with unknown functions remain (Rhee & Mutwil, 2014). The establishment of model plants such as *A. thaliana* has played a substantial role in the development of comparative genomics, a field which uses genome sequence homology as a means of identifying gene function in previously uncharacterized organisms (Smith et al., 2019). Although comparative genomics is a useful tool for identifying gene homologs and syntenic regions of genomes, it becomes less practical when genes of interest are part of genus-specific specialized metabolic pathways that lack clear homologs in other plants. Specialized metabolites have become a key area of study in pharmacology, agriculture, and cosmetics, and so there exists a significant need for new techniques that would better allow for the identification and characterization of these divergently evolved pathways.

Sequencing plant genomes has revealed new information about the genomic organization of metabolic genes. While it was previously thought that eukaryotic biosynthetic genes were distributed non-continuously across distinct chromosomes, recent findings suggest that plant genes associated with specialized metabolic pathways may physically aggregate within the genome in biosynthetic gene clusters similar to a bacterial operon (Nützmann, Huang, & Osbourn, 2016). In plants, biosynthetic gene clustering is hypothesized to allow for the coinheritance of entire specialized pathways, thereby decreasing the likelihood of incomplete pathway inheritance that could lead to toxic intermediate build-up (Kim & Buell, 2015). Mining biosynthetic gene clusters can be useful for identifying specialized metabolic genes that contribute to the same pathway (Nützmann et al., 2016). The online platform plantiSMASH can be used to identify genomic

loci that resemble biosynthetic gene clusters and can also use transcriptomics data to prioritize candidates based on coexpression (Kautsar, Suarez Duran, & Medema, 2018). In the absence of biosynthetic gene clusters, candidate gene identification relies on transcriptomic or metabolic approaches that require gene expression induction, which can lead to larger pools of candidate genes. While there have been many recent advances in biosynthetic gene cluster identification, the products of many of these metabolic clusters have not yet been elucidated (Polturak, Liu, & Osbourn, 2022).

One additional technique for identifying candidate metabolic genes is by correlating tissue-level expression with metabolite presence and abundance. This method is especially useful in plants, which often synthesize and compartmentalize defense metabolites to ensure that they reach their target efficiently without inducing autotoxicity (Delli-Ponti, Shivhare, & Mutwil, 2021). In order to correlate gene expression with metabolite presence, it is important to first identify the tissue, cell type, or developmental stage where the specialized metabolite is synthesized. This can be done using mass spectrometry, where metabolite(s) of interest are detected using mass spectrometry and their relative abundance is compared across samples (Saito & Matsuda, 2010). After generating metabolomics data, transcriptomics can be used to identify genes that are differentially expressed across the same tissues, cell types, or developmental stages that were used for collecting metabolomics data. Correlating gene expression with metabolite presence and abundance can be very effective in identifying genes in biosynthetic pathways. This approach is less effective when the specialized metabolites are not synthesized in the tissue in which they are localized. For example, the plant defensive compound nicotine is synthesized in the roots of *Nicotiana tabacum* (tobacco) and transported to the leaves, where it serves as an insecticide (Baldwin, 1989). Although this approach has its limitations, it is a useful technique for identifying candidate primary and secondary metabolic genes.

In addition to forming biosynthetic gene clusters, specialized metabolic genes are often coregulated by a common set of transcription factors and consequently, coexpressed. One method for identifying these expression patterns is to use global coexpression network analysis, where genes, which are represented by nodes, are linked together based on overlapping expression profiles to form modules (Wisecaver et al., 2017). Though this approach offers a high-throughput method for identifying candidate specialized metabolic pathways, it is important to recognize that many specialized pathway genes within a pathway are not necessarily coexpressed. Additionally, it can be difficult to delineate separate metabolic pathways that may be

expressed in response to the same elicitors. Despite these limitations, global coexpression network analysis remains a powerful tool for identifying candidate secondary metabolic genes in plants.

While the above-mentioned computational techniques have been instrumental in predicting gene function and for biosynthetic pathway discovery, these approaches may still yield long lists of candidate genes that all need to be screened using in vivo or in vitro enzyme assays. To narrow down lists of candidate enzymes further, computational substrate docking using virtual screens is a protein-based approach that may be used to predict enzyme function. Additionally, virtual screening may be applied broadly to studying metabolic enzymes that are targets for herbicides, promiscuous enzymes, or enzymes of unknown function.

1.2 Docking and virtual screening as a computational tool for functional prediction of plant metabolic enzyme activity

Docking is a commonly used computational tool used to model interactions between a three-dimensional protein structure and a small molecule. Similarly, virtual screening is a docking approach that iteratively docks a library of molecules with a target protein. For decades, protein biochemists, molecular biologists, and pharmaceutical industries have been using protein structure determination or protein homology modeling coupled with small-molecule docking to investigate molecular interactions. This line of research has been made possible by the availability of open-source docking software that does not require advanced programming knowledge (Villoutreix et al., 2007). One of the most widely used docking programs, AutoDock Vina, uses a simple scoring function to efficiently evaluate intermolecular interactions within a given protein–ligand complex and outputs a prediction of protein–ligand binding affinities and binding conformations, which can be modeled in three-dimensional visualization programs such as PyMol (Trott & Olson, 2010). Docking is an excellent tool for identifying candidate substrates, cofactors, and regulators for an enzyme of interest, and in plant biology research, it has primarily been used to model interactions between a single protein and one metabolite. Examples of instances where docking has been used to study plant enzymes include: the *Arabidopsis* GH3.15 that conjugates amino acids to the auxinic hormone indole-3-butyric acid; a noroxomaritidine reductase involved in alkaloid biosynthesis in daffodils (*Narcissus* spp.); and a rice naringenin *O*-methyltransferase involved in phytoalexin synthesis (Kilgore, Holland, Jez, & Kutchan, 2016; Murata et al., 2020; Sherp, Westfall, Alvarez, & Jez, 2018).

In the past, protein-based methods for narrowing down lists of candidate enzymes in biosynthetic pathways have been limited by the availability of reliable protein structures or structural models. Though cryo-electron microscopy, nuclear magnetic resonance (NMR), and X-ray crystallography have been instrumental for solving protein structures, each of these approaches has their challenges. The development of web-based programs with simple interfaces for homology-based modeling, such as SWISS-MODEL and Phyre2, has allowed for the generation of highly accurate three-dimensional protein structures (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015; Waterhouse et al., 2018). However, proteins that lack known homologs, as in the case of many secondary metabolic enzymes, may be prone to structural inaccuracy. Recent advances in AI-based structure prediction with the release of AlphaFold has substantially improved our ability to predict the three-dimensional structure of enzymes from any species (Jumper et al., 2021). AlphaFold is a machine learning program that integrates knowledge of biophysical dynamics with protein evolutionary history analysis to generate highly accurate structural predictions, even if no structural homolog is known. Being able to generate a structural model of any protein of interest has the potential to improve our ability to study specialized metabolism and decipher the molecular underpinnings of plant metabolism broadly.

In medicinal chemistry and pharmacology, it is common practice to screen compound–enzyme interactions computationally before proceeding with empirical experiments, especially when there are hundreds or thousands of compounds and several target proteins (Rester, 2008). This same approach can be used for investigating plant metabolism and aid in identifying candidate enzymes in biosynthetic pathways. Virtual screens can be implemented for all metabolic enzymes, including glycosyltransferases, oxygenases (i.e., cytochrome P450s), oxidoreductases, ligases, hydrolases, or terpene synthases. Aside from identifying enzyme substrates, virtual screening may be used to investigate many open questions in plant metabolism, including understanding the substrate promiscuity of enzymes or identifying competitive and allosteric inhibitors of an enzyme. The structure-based virtual screening methods described here use open access programs that only require a local PC or Mac computer. Because AutoDock Vina is open-source, fast, and has a wealth of online tutorials, user manuals, and discussion forums available to support new users, the protocol will focus on the use of this program (Eberhardt, Santos-Martins, Tillack, & Forli, 2021; Trott & Olson, 2010). While having a working understanding of protein structural visualization programs such

as PyMOL or Chimera would be helpful, it is not necessary, and resources such as PyMOL Wiki are available online.

2. Designing a virtual screen experiment

Before conducting a virtual screen, the enzyme of interest and the library of compounds that will be screened will need to be prepared. While there are several programs available that will run a virtual screen, the information below will focus on preparing compounds and enzymes for docking using the widely used and freely available program AutoDock Vina (Trott & Olson, 2010), and the files will be prepared for docking using AutoDock Tools, a graphical user interface that is part of the MGLTools software suite. Because these programs are popular for docking, many online resources and published protocols are readily available (Forli et al., 2016).

To demonstrate how virtual screens are executed, the protocols below will use a glycosyltransferase from the model plant *A. thaliana,* UGT74F2 (AT2G43820), that has been functionally characterized as a UDP-dependent glycosyltransferase (UGT) that glycosylates the carboxylate of the plant hormone salicylic acid (SA), forming an SA glucose ester (SGE) (Lim et al., 2002). This enzyme is known to be promiscuous and can also glycosylate the hydroxyl of SA (forming SA 2-O-beta-D-glucose; SAG), as well as other benzoate substrates, including the tryptophan pathway intermediate anthranilate (2-aminobenzoate), benzoic acid, and 3-hydroxybenzoic acid (George Thompson, Iancu, Neet, Dean, & Choe, 2017; Lim et al., 2002; Quiel & Bender, 2003). UGTs use nucleotide-activated sugars as sugar donors in the transferase reaction, and UGT74F2, as well as other plant UGTs, uses UDP-glucose as the sugar donor (Akere et al., 2020). These UGTs have a variable N-terminal domain and a C-terminal nucleotide-sugar binding domain that contains a conserved 44 amino acid motif known as a Plant Secondary Product GT box, and substrates bind UGTs in a cleft between these two domains.

To prepare ligands and protein input files for virtual screening, you will need to begin by downloading and installing MGLTools from: http://mgltools.scripps.edu/downloads. On a Mac, users will also need to download X11 from http://xquartz.org in order to run AutoDock Tools.

2.1 Generating a compound library

The first step in conducting a virtual screen is to decide which compounds to include in the screen given the enzyme of interest. While several online

databases of ligand files exist, the compounds in these databases are primarily targeted for human health and medicine, including ZINC Docking (Irwin et al., 2020; Sterling & Irwin, 2015), PubChem (Kim et al., 2016), and ChEMBL (Bento et al., 2014). However, because numerous plant metabolites are used as pharmaceuticals or nutraceuticals, ready-to-dock molecules for many commonly investigated plant-produced compounds are available for download in a ".mol2" file format from ZINC Docking, including intermediates in primary metabolism, plant hormones, specialized metabolites, and herbicides (Irwin et al., 2020; Sterling & Irwin, 2015). While by no means comprehensive, a list of plant metabolites that are available for download from ZINC Docking has been included in Tables 1 and 2. Aside from the hormones, metabolites, and herbicides listed in Tables 1–3, primary

Table 1 List of plant hormones and hormone-related metabolites included in the compound library ranked by molecular mass.

Metabolite	Formula	Molecular mass (g/mol)
Phenylacetic acid	$C_8H_8O_2$	136.15
Salicylic acid	$C_7H_6O_3$	138.12
Indole-3-acetic acid (IAA)	$C_{10}H_9NO_2$	175.18
Indole-3-butyric acid	$C_{12}H_{13}NO_2$	203.24
4-Chloroindole-3-acetic acid	$C_{10}H_8ClNO_2$	209.63
Jasmonic acid	$C_{12}H_{18}O_3$	210.27
Kinetin	$C_{10}H_9N_5O$	215.21
cis-Zeatin	$C_{10}H_{13}N_5O$	219.24
6-Benzylaminopurine	$C_{12}H_{11}N_5$	225.25
Abscisic acid	$C_{15}H_{20}O_4$	264.32
IAA-glutamine	$C_{15}H_{17}N_3O_4$	303.31
Sorgolactone	$C_{18}H_{20}O_5$	316.30
Gibberellic acid	$C_{19}H_{22}O_6$	346.37
Strigol	$C_{19}H_{22}O_6$	346.40
Orobanchol	$C_{19}H_{22}O_6$	346.40
Kinetin riboside	$C_{15}H_{17}N_5O_5$	347.33
Kinetin-9-N-glucoside	$C_{16}H_{19}N_5O_6$	377.35

Table 2 List of plant metabolites included in the compound library ranked by general classification and molecular mass.

Specialized metabolite	Description	Formula	Molecular mass (g/mol)
Nicotine	Alkaloid	$C_{10}H_{14}N_2$	162.23
Caffeine	Alkaloid	$C_8H_{10}N_4O_2$	194.19
Camalexin	Alkaloid	$C_{11}H_8N_2S$	200.26
Morphine	Alkaloid	$C_{17}H_{19}NO_3$	285.34
Galantamine	Alkaloid	$C_{17}H_{21}NO_3$	287.35
Capsaicin	Alkaloid	$C_{18}H_{27}NO_3$	305.41
Quinine	Alkaloid	$C_{20}H_{24}N_2O_2$	324.4
Berberine	Alkaloid	$C_{20}H_{18}NO_4^+$	336.4
Strictosidine	Alkaloid	$C_{27}H_{34}N_2O_9$	530.57
Vinblastine	Alkaloid	$C_{46}H_{58}N_4O_9$	811
Benzaldehyde	Aromatic	C_7H_6O	106.12
Benzoate	Aromatic	$C_7H_5O_2^-$	121.11
Cinnamaldehyde	Aromatic	C_9H_8O	132.16
p-Coumaryl alcohol	Aromatic	$C_9H_{10}O_2$	150.17
Methyl anthranilate	Aromatic	$C_8H_9NO_2$	151.17
Vanillin	Aromatic	$C_8H_8O_3$	152.15
Methyl salicylate	Aromatic	$C_8H_8O_3$	152.15
Gallic acid	Aromatic	$C_7H_6O_5$	170.12
Ferulic acid	Aromatic	$C_{10}H_{10}O_4$	194.18
Anthraquinone	Aromatic	$C_{14}H_8O_2$	208.21
DIMBOA	Aromatic	$C_9H_9NO_5$	211.17
Resveratrol	Aromatic	$C_{14}H_{12}O_3$	228.24
Catechin	Aromatic	$C_{15}H_{14}O_6$	290.26
Quercetin	Aromatic	$C_{15}H_{10}O_7$	302.23
Dhurrin	Aromatic	$C_{14}H_{17}NO_7$	311.29
Tetrahydrocannabinol	Aromatic	$C_{21}H_{30}O_2$	314.45
Bergamottin	Aromatic	$C_{21}H_{22}O_4$	338.4

Continued

Table 2 List of plant metabolites included in the compound library ranked by general classification and molecular mass.—cont'd

Specialized metabolite	Description	Formula	Molecular mass (g/mol)
Rosmarinic acid	Aromatic	$C_{18}H_{16}O_8$	360.3
Podophyllotoxin	Aromatic	$C_{22}H_{22}O_8$	414.41
Allicin	Sulfur-containing	$C_6H_{10}OS_2$	162.28
Glucobrassicin	Sulfur-containing	$C_{16}H_{19}N_2O_9S_2^-$	447.46
Limonene	Terpene	$C_{10}H_{16}$	136.24
Linalool	Terpene	$C_{10}H_{18}O$	154.25
Campesterol	Terpene	$C_{28}H_{48}O$	400.68
Betulinic acid	Terpene	$C_{30}H_{48}O_3$	456.7
Lycopene	Terpene	$C_{40}H_{56}$	536.87
Beta-carotene	Terpene	$C_{40}H_{56}$	536.87

Aromatics refers to metabolites that contain an aromatic ring (i.e., flavonoids, polyphenols, coumarins, etc.), and terpenes include mono-, di-, and triterpenes.

Table 3 List of herbicides included in the compound library listed alphabetically.

Herbicide	Formula	Molecular mass (g/mol)
Aatrex	$C_8H_{14}ClN_5$	215.69
Bentazon	$C_{10}H_{12}N_2O_3S$	240.28
Caprylic acid	$C_8H_{16}O_2$	144.21
Clethodim	$C_{17}H_{26}ClNO_3S$	359.9
Clomazone	$C_{12}H_{14}ClNO_2$	239.7
Clopyralid	$C_6H_3Cl_2NO_2$	192
Cycloate	$C_{11}H_{21}NOS$	215.36
Ethalfluralin	$C_{13}H_{14}F_3N_3O_4$	333.26
Ethofumesate	$C_{13}H_{18}O_5S$	286.34
Fluazifop	$C_{19}H_{20}F_3NO_4$	327.25
Flumioxazin	$C_{19}H_{15}FN_2O_4$	354.1

Table 3 List of herbicides included in the compound library listed alphabetically.—cont'd

Herbicide	Formula	Molecular mass (g/mol)
Fomesafen	$C_{15}H_{10}ClF_3N_2O_6S$	460.7
Glyphosate	$C_3H_8NO_5P$	169.07
Halosulfuron-methyl	$C_{13}H_{15}ClN_6O_7S$	434.81
Imazethapyr	$C_{15}H_{19}N_3O_3$	289.33
Linuron	$C_9H_{10}Cl_2N_2O_2$	249.1
Metam sodium	$C_2H_4NNaS_2$	129.18
Mesotrione	$C_{14}H_{13}NO_7S$	339.32
Metribuzin	$C_8H_{14}N_4OS$	214.29
Napropamide	$C_{17}H_{21}NO_2$	271.16
Nicosulfuron	$C_{15}H_{18}N_6O_6S$	410.4
Norflurazon	$C_{12}H_9ClF_3N_3O$	303.04
Oxyfluorfen	$C_{15}H_{11}ClF_3NO_4$	361.7
Paraquat	$C_{12}H_{14}Cl_2N_2$	257.16
Pelargonic acid	$C_9H_{18}O_2$	158.23
Pendimethalin	$C_{13}H_{19}N_3O_4$	281.31
Phenmedipham	$C_{16}H_{16}N_2O_4$	300.31
Prometryn	$C_{10}H_{19}N_5S$	241.36
Pronamide	$C_{12}H_{11}Cl_2NO$	256.12
Pyraflufen-ethyl	$C_{15}H_{13}Cl_2F_3N_2O_4$	413.2
Pyroxasulfone	$C_{12}H_{14}F_5N_3O_4S$	391.06
Rimsulfuron	$C_{14}H_{17}N_5O_7S_2$	431.4
Saflufenacil	$C_{17}H_{17}ClF_4N_4O_5S$	500.9
Sethoxydim	$C_{17}H_{29}NO_3S$	327.5
Simazine	$C_7H_{12}ClN_5$	201.66
Tembotrione	$C_{17}H_{16}ClF_3O_6S$	440.8
Terbacil	$C_9H_{13}ClN_2O_2$	216.67
Trifluralin	$C_{13}H_{16}F_3N_3O_4$	335.28

metabolites like amino acids, nucleotides, and intermediates in carbohydrate and lipid metabolism that are conserved across domains of life are also available from ZINC Docking and may be useful to include in a virtual screening experiment.

In addition to the compounds included in Tables 1–3, the example virtual screen involving the Arabidopsis UGT74F2 will also include: 3-hydroxybenzoic acid, 4-hydroxybenzoic acid, 2,3-dihydroxybenzoic acid, 2,4-dihydroxybenzoic acid, 2,5-dihydroxybenzoic acid, 2,6-dihydroxybenzoic acid, 3,4-dihydroxybenzoic acid, 3-hydroxyanthranilate, chorismate, and tryptophan. The ".mol2" files for each of these ligands were downloaded from ZINC Docking.

While online databases of metabolites have many of the common plant metabolites, intermediates in specialized metabolism are not likely to be included. If a compound of interest is not available to download, as a flexible 3D formatted file, structural files can be interconverted into acceptable file formats, like ".mol2", using the open-source tool Open Babel (O'Boyle et al., 2011).

2.2 Preparing molecules for a virtual screen

1. The coordinate files for the ligands from Section 2.1 will now need to be converted to PDBQT (Protein Data Bank, Partial Charge (Q), & Atom Type (T)) files for downstream applications. The files can be opened as ligands in AutoDock Tools and converted to the ".pdbqt" file extension (Forli et al., 2016). To do this, click on "Ligand" in the toolbar and select "input" to open each of the ".mol2" files. To convert them to ".pdbqt" files, click on "Ligand" and select "output" and save the file in a directory where it can be easily located. The name for each ligand file should not contain spaces
2. Once the ligand files have been generated, save them in a folder named "bin" inside a "Vina" folder on your computer's desktop. Alternatively, the files may be renamed with an abbreviation or a number, and a spreadsheet could be used to connect the full name of the compound to the abbreviated file name
3. Using a text editor, create a file named Ligands.txt. Type the name of each ligand file (i.e., benzoate.pdbqt). Each line in this file should contain only a single ligand file name. This file should be saved in Desktop/Vina/bin. NOTE: If the spelling of the ligand file name does not exactly match the spelling in the Ligands.txt file, downstream codes will not run

4. To prepare the enzyme of interest, you will need to obtain structural coordinate of the protein. Several options exist for finding or generating a coordinate file for the enzyme:
   a. If a solved structure of the enzyme of interest is available, the ".pdb" file can be downloaded from the Protein Data Bank (rcsb.org). If the protein structure was solved in complex with a ligand that occupies the binding site (i.e., active site or allosteric site), then the ".pdb" file can be opened using a text editor, and the three-dimensional coordinates of the ligand can be manually deleted from the file. To determine that the ligand has been successfully removed, the modified ".pdb" file can be opened using a structural visualization program like PyMOL and visually inspected. Extraneous ions, water molecules, ligands, and cofactors can also be deleted using a text editor. Solved structures may also have multiple biological assemblies (e.g., dimers) within one asymmetric unit. While not necessary, additional biological assemblies may be deleted manually using a text editor and inspected using PyMOL
   b. As of this writing, the proteomes for the model plants *A. thaliana, Zea mays, Glycine* max, and *Oryza sativa* are available for download online from the AlphaFold Protein Structure Database (www.alpha fold.ebi.ac.uk; Jumper et al., 2021). For example, searching for "*Arabidopsis thaliana* UGT74F2" returns an entry for this protein, and a structural model can be downloaded as a ".pdb" file. Note that these models do not contain cofactors, metal ions, water molecules, or ligands. It may be important to add cofactors to the structure using a single docking run (see the note at the end of this section).
   c. For non-model plants or for proteins that are not available from the AlphaFold database, protein models can be generated from online servers, such as SWISS-MODEL (swissmodel.expasy.org) or Phyre2 (www.sbg.bio.ic.ac.uk/~phyre2), using an amino acid sequence for the enzyme of interest
5. Once a coordinate file for the enzyme(s) of interest has been obtained, the file can be prepared using AutoDock Tools. The protein used here, UGT74F2 from *A. thaliana*, has been previously crystalized (PDB ID: 5U6M; George Thompson et al., 2017), and the file was first prepared by removing the salicylic acid ligands from the dimer. The UDP that was cocrystalized with the enzyme was left in the file because it is the product of the reaction after the activated nucleotide sugar (i.e., UDP-glucose) transfers the sugar onto a nucleophilic acceptor substrate. Alternatively,

the AlphaFold-generated structural coordinates may have been used, but the protein structure would have been apo, meaning that UDP would have to be docked into the structure. To open the ".pdb" coordinate file for the enzyme, use the topmost toolbar and select "File" and then "Read Molecule." Protein structures typically do not include hydrogens, so these may be added to the enzyme using AutoDock Tools by clicking "Edit" in the toolbar and then "Hydrogens" (Forli et al., 2016). Hydrogens should appear on the enzyme in the protein viewing window. To export the protein as a ".pdbqt" file, click "Grid," then "Macromolecule" and "Choose" and select the protein. A window will pop up so the file can be saved in the Desktop/Vina/bin folder with a ".pdbqt" extension

6. To define the docking area that encompasses the active site or binding pocket of interest, click "Grid" and then "Grid Box." This is the area in which Vina will search for binding interactions between the enzyme and each ligand, so it is important that the space is large enough to cover the entire area of interest and allow the ligands to rotate freely but not so large that the molecules are predicted to bind the protein superficially. The size and center of the search space in x, y, and z dimensions of the grid box, can be altered so the entire search space is encompassed by the coordinates. It is also important to rotate the protein in the AutoDock Tools viewer to see that the binding pocket of interest is covered in all dimensions
7. Once the grid box dimensions have been determined, these values will be recorded in a text file saved in the Desktop/Vina/bin subfolder. This plain text file can be made using a plain text editor, like TextEdit or Notepad

```
conf_vs.txt

receptor = UGT74F2.pdbqt

center_x = 46.7
center_y = 76.1
center_z = 81.6

size_x = 40
size_y = 40
size_z = 40

exhaustiveness = 8
num_modes = 10
```

The name of the enzyme coordinate file is listed as "receptor =" and the center x, y, and z coordinates were determined from AutoDock

Tools, as was the size of the box in x, y, and z dimensions. The exhaustiveness of 8 is default and is the number of Monte-Carlo iterated searches. The number of modes (num_modes) is the maximum number of ligand conformations that will be generated per ligand. Increasing this number may increase the total time that it takes to run Vina.

8. This process was repeated to generate a structural model of UGT74F2 that had the activated nucleotide sugar donor UDP-glucose bound in the active site. To do this, the coordinates for UDP were removed from the 5U6M PDB file above using a text editor. The apo enzyme structure was used for a single docking run using AutoDock Vina, and the coordinates of the UDP-glucose conformation that had the highest binding affinity and bound in a logical conformation was exported as a PDB file and pasted into the text file of the apo enzyme model and saved as a new UDP-glucose bound model. This new combined file was opened using PyMOL to ensure that the structural model had UDP-glucose in the active site. This file was then opened in AutoDock Tools for conversion to a ".pdbqt" file and to determine the grid box dimensions

Note: If an important cofactor for the protein of interest—such as heme in cytochrome P450 monooxygenases, NADH or NADPH in dehydrogenases, or S-adenosyl methionine in SAM-dependent methyltransferases—is not part of the structural file, a single docking run using AutoDock Vina could be used to place the cofactor in the active site. The coordinate file for cofactors may be obtained from the Protein Data Bank, converted to a ".pdbqt" file, and docked into the active site. Coordinates for the best docking confirmation of the cofactor could be pasted into the ".pdb" file of the protein above "END" using a text editor.

3. Virtual screening using AutoDock Vina

The virtual screen presented here uses AutoDock Vina, which uses a rapid gradient-optimization conformational search (Forli et al., 2016). The scoring function estimates the force of non-covalent interactions between compounds and the target protein using mathematical models, and the user does not need to have working knowledge of these models in order to use the program. This straightforward approach is easily implemented by molecular biologists, and the length of time needed for the screen to run depends on the factors such as the number of ligands included in the virtual screen, as well as the number of modes for each ligand.

3.1 Software needed

Windows users will need to download Perl to run the virtual screen script from https://www.perl.org/get.html. Perl is installed on MacOS by default. AutoDock Vina can be downloaded online at http://vina.scripps.edu. The following protocol uses version 1.1.2 for its ease of use. Move the two Unix executable files (titled vina and vina_split) from the Autodock Vina folder that was downloaded into the Desktop/Vina/bin subfolder. These are needed to ensure the following steps proceed properly.

3.2 Script for running the virtual screen

A Perl script enables iterative docking runs with AutoDock Vina using a Mac or Windows system. This file should be added to the Desktop/Vina/bin folder. The file can be created using a text editor and should contain the following information:

Vina_mac.pl

```
#!/usr/bin/perl
print"Ligand_file:\t";
$ligfile=<STDIN>;
chomp $ligfile;
open (FH,$ligfile)||die "Cannot open file\n";
@arr_file=<FH>;

for($i=0;$i<@arr_file;$i++)
{
print"@arr_file[$i]\n";
@name=split(/\./,@arr_file[$i]);
}
for($i=0;$i<@arr_file;$i++)
{
    chomp @arr_file[$i];
    print"@arr_file[$i]\n";
    system("./vina –config conf_vs.txt –ligand @arr_file[$i] –log
@arr_file[$i]_log.log");
    }
```

Modification to Perl script for Microsoft OS:

```
    system("vina.exe –config conf_vs.txt –ligand @arr_file[$i] –log
@arr_file[$i]_log.log");
```

3.3 Procedure on a Mac OS

1. Before starting the run, check that the following files are in the Desktop/Vina/bin folder:

a. conf_vs.txt
b. Ligands.txt
c. Vina_mac.pl (or Vina_windows.pl)
d. vina
e. vina_split
f. UGT74F2.pdbqt
g. ".pdbqt" files for each ligand

2. Open Terminal (Mac) or Command Prompt (Windows), and type:
cd Desktop/Vina/bin
perl Vina_mac.pl (OR **perl Vina_windows.pl**)

When you hit "Enter," you will be prompted to type in the name of the ligand file. Type: **Ligands.txt** and hit "Enter." After this, Vina will start running. Here is an example of the Terminal log during one compound docking run:

```
cinnamaldehyde.pdbqt
WARNING: The search space volume > 27000 Angstrom^3 (See FAQ)
Output will be cinnamaldehyde_out.pdbqt
Detected 4 CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: 1584341810
Performing search ...
0%  10  20  30  40  50  60  70  80  90  100%
|---|---|---|---|---|---|---|---|---|---|
**********************************************
done.
Refining results ... done.
mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1          -5.9      0.000      0.000
   2          -5.6      9.406     10.829
   3          -5.4      2.469      4.623
   4          -5.1      8.247     10.127
   5          -5.1     12.839     14.177
   6          -5.0      9.418     11.085
   7          -4.9      9.755     10.414
   8          -4.9     12.924     14.300
   9          -4.9      3.682      5.598
  10          -4.8     14.032     15.374
Writing output ... done.
```

The length of the run will vary based on the number of ligand files and the number of modes. Do not close the Terminal window during this time.

3. When your run is complete, all of the coordinate files for each of the docked ligands (i.e., ligand_out.pdbqt), as well as the output log files that include the binding affinities for each conformation, will be in the Desktop/Vina/bin folder

Note: Before running a new virtual screen, it is important to move the ".log" and "..._out.pdbqt" files from the previous run to a new folder so they are not overwritten.

4. Interpreting and visualizing screening results

To interpret the results, the protein can be opened in a protein visualization program, like PyMOL or UCSF Chimera. PyMOL can be downloaded from https://pymol.org/2/. In the same window, each of the ligand files can be opened and visualized in relation to the active site or a specified cavity on the enzyme of interest. Because the conf_vs.txt file used a num_modes setting of 10, a maximum of 10 conformations were generated for each compound. These conformations should be manually inspected to determine the conformation that is most logical (i.e., oriented within the active site and not on the surface of the protein) and has the lowest computational binding affinity (i.e., a large negative number) in kcal/mol. The conformations are shown in increasing order of binding affinity, and while the first conformation has the highest binding affinity, it does not always bind within the active site or in an orientation that would make sense for catalysis. If many of the molecules are binding on the surface of the protein, the search space dimensions can be redrawn, and the screen can be repeated with a more refined search window. If the grid box dimensions are stringent and the compound does not bind within those constraints, the results may have fewer than 10 conformations for some compounds.

For metabolic enzymes, one way to assess the accuracy of ligand binding is to use the measurement feature in PyMOL to determine the distance between the compound and the cofactor (Fig. 1). In the toolbar of PyMOL, click on "Wizard" and select "Measurement." This allows for the distance between individual atoms to be determined. Click on an atom on the ligand that is acted upon by the enzyme (e.g., the predicted ligand hydroxyl that is glycosylated by the enzyme), and then click on a second atom elsewhere in the enzyme like a catalytic residue (if known) or an atom

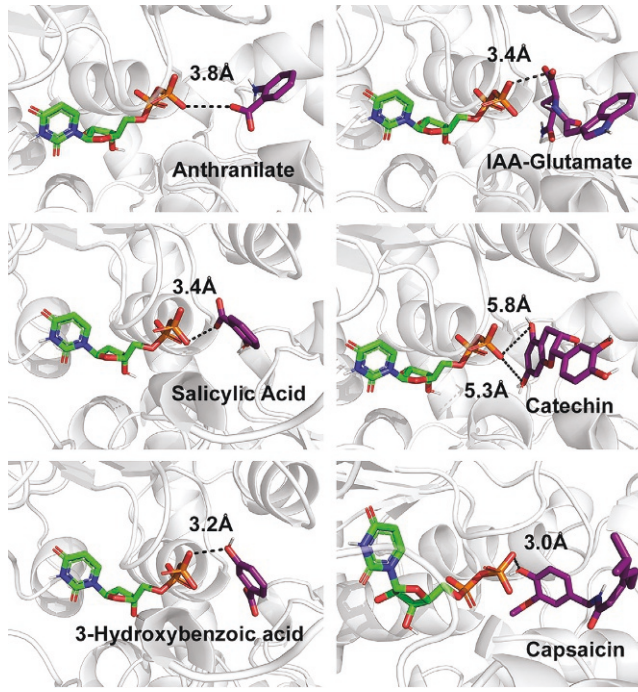

Fig. 1 Measured distances between functional groups on docked ligands and the nearest phosphate oxygen on UDP in the AtUGT74F2 active site. In the three panels on the left, three compounds (anthranilate, salicylic acid, and 3-hydroxybenzoic acid) that are confirmed substrates for the enzyme are shown, while compounds that have a high binding affinity but unknown activity (IAA-glutamate, catechin, and capsaicin) are shown in the three panels on the right.

on the cofactor. The distance will be displayed as a label in Angstroms (Å) above a dashed line that connects the two atoms that were selected. Once you are finished measuring distances between atoms, click "Done" under "Measurement" in the right-hand panel.

While inspecting each ligand, note the conformation(s) that have a low binding affinity and accurately bind within the active site. Vina has calculated the binding affinity for each mode, and the values can be found in the output ".log" files that were generated for each ligand. Taken together, the binding affinities for the ligand conformations that were noted by visual inspection may be added to a spreadsheet (Table 4 and 5). When the results have been tabulated, the compounds can be ranked by their binding affinity values from lowest to highest. The results from the virtual screens of UGT74F2 that had either UDP-glucose or UDP bound in the active site are included in Tables 4 and 5.

Table 4 Ranking of binding affinities from the *Arabidopsis thaliana* UGT74F2 virtual screen with UDP-glucose bound in the active site.

Compound	Binding mode	Binding affinity (kcal/mol)
Capsaicin	1	−7.2
2,4-Dihydroxybenzoic acid[a]	1	−6.5
2,3-Dihydroxybenzoic acid[a]	1	−6.5
Gallic acid	3	−6.1
3-Hydroxyanthranilate	3	−6.1
3,4-Dihydroxybenzoic acid[a]	5	−6.1
Salicylic acid[b]	3	−6
Nicotine	1	−6
Anthranilate[b]	2	−5.9
3-Hydroxybenzoic acid[b]	3	−5.9
4-Hydroxybenzoic acid[a]	3	−5.9
2,5-Dihydroxybenzoic acid[a]	5	−5.8
Phenylacetic acid	2	−5.7
p-Coumaryl alcohol	2	−5.6
Indole-3-butyric acid	7	−5.6
Benzoic acid[b]	3	−5.6
Clopyralid	4	−5.4
2,6-Dihydroxybenzoic acid[a]	3	−5
Glyphosate	4	−4.4
Metam	2	−2.8

[a]No activity detected in an in vitro assay.
[b]Activity has been confirmed in vitro.
Binding affinities are listed for the binding mode that was closest to the UDP-glucose cosubstrate (measured in angstroms). The distance represents the approximate distance between the nearest atom that could participate in the glycosyl transfer reaction (e.g., hydroxyl oxygen) and the hydroxyl on carbon-6 of glucose. Compounds with a predicted distance of >5 Å from the oxygen atoms of the gamma phosphate were not included in the table. Compounds are ranked from highest binding affinity to lowest binding affinity as determined by AutoDock Vina, and molecules that bound in the active site but did not have a functional group that could be glycosylated were not included in the table.

Table 5 Ranking of binding affinities from the *Arabidopsis thaliana* UGT74F2 virtual screen with UDP bound in the active site.

Compound	Binding mode	Binding affinity (kcal/mol)
IAA-glutamate	1	−8.8
Catechin	1	−8.8
Tryptophan	1	−7.8
Resveratrol	1	−7.4
6-Benzylaminopurine	1	−7.2
DIMBOA	1	−7.1
Linuron	2	−7.1
Kinetin-riboside	8	−7
Phenmedipham	3	−7
Dhurrin	6	−6.9
cis-Zeatin	3	−6.8
IAA	2	−6.8
Indole-3-butyric acid	2	−6.8
4-Chloroindole-3-acetic acid	1	−6.8
Capsaicin	4	−6.4
Ferulic acid	2	−6.4
Bentazon	5	−6.3
2,3-Dihydroxybenzoic acid[a]	4	−6.2
2,5-Dihydroxybenzoic acid[a]	4	−6.2
Caffeine	4	−6.2
2,6-Dihydroxybenzoic acid[a]	3	−6.1
2,4-Dihydroxybenzoic acid[a]	1	−6.1
3-Hydroxyanthranilate	6	−5.9
Anthranilate[b]	2	−5.9
Nicotine	3	−5.8
Terbacil	5	−5.8
Clopyralid	2	−5.8

Continued

Table 5 Ranking of binding affinities from the *Arabidopsis thaliana* UGT74F2 virtual screen with UDP bound in the active site.—cont'd

Compound	Binding mode	Binding affinity (kcal/mol)
p-Coumaryl alcohol	2	−5.7
Salicylic acid[b]	5	−5.6
Phenylacetic acid	4	−5.6
3-Hydroxybenzoic acid[b]	8	−5.5
Paraquat	3	−5.5
Simazine	3	−5.4
Methyl anthranilate	8	−5.3
4-Hydroxybenzoic acid[a]	8	−5.3
Methyl salicylate	10	−5
Glyphosate	5	−5
Caprylic acid	1	−4.9
Nonoate	2	−4.8
Metam	1	−3.1

[a]No activity detected in an in vitro assay.
[b]Activity has been confirmed in vitro.
Binding affinities are listed for the binding mode that was closest to the UDP product (measured in angstroms). The distance represents the approximate distance between the nearest atom that could participate in the glycosyl transfer reaction (e.g., hydroxyl oxygen) and the oxygen atoms on the beta-phosphate of UDP. Compounds with a predicted distance of greater than 6.0 Å from the terminal phosphate oxygen atoms were not included in the table. Compounds are ranked from highest binding affinity to lowest binding affinity as determined by AutoDock Vina, and molecules that bound in the active site but did not have a functional group that could be glycosylated were not included in the table.

For UGT enzymes, analyzing the results of a virtual screen with both the UDP-glucose cosubstrate as well as the UDP product bound is informative since the glucose moiety is large and occupies a portion of the active site (Figs. 1 and 2). When UDP-glucose is bound, larger molecules may be precluded from binding the active site pocket, especially if the pocket is small. However, a virtual screen that has UDP bound in the active site may allow larger or glycosylated molecules to bind. Taken together, the UDP-glucose bound structure is useful for identifying putative substrates for this enzyme, while the UDP-bound structure may inform putative products of the reactions. In either docking simulation, putative small-molecule regulators (i.e., competitive inhibitors) may also be identified.

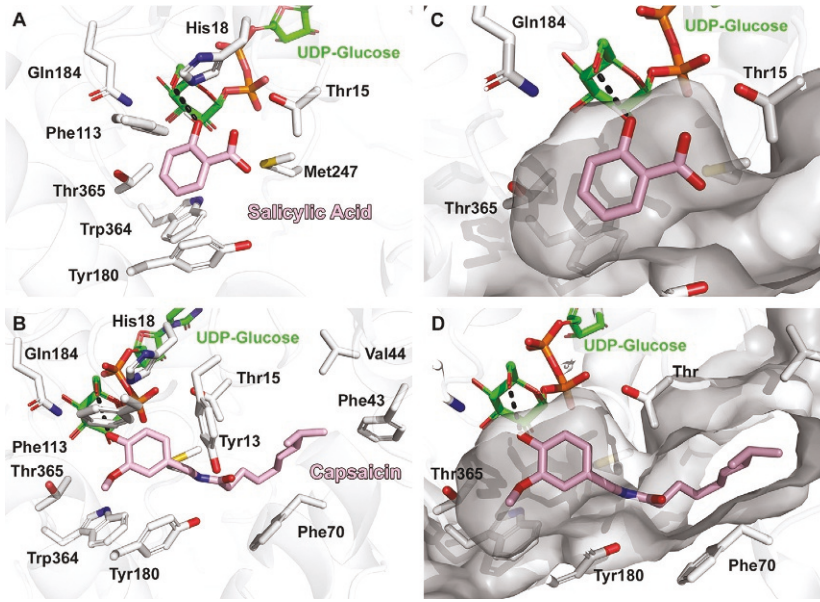

Fig. 2 Docking results for AtUGT74F2 with UDP-glucose in the active site. A. View of the active site with salicylic acid docked. Dashed lines are used to show how the distance between the hydroxyl of salicylic acid and the hydroxyl on carbon-6 of glucose was measured. Active site residues are displayed as sticks. B. View of the active site with capsaicin docked, including hydrophobic residues distal from the site of glycosyl transfer. Dashed lines are used to show how the distance between the hydroxyl of salicylic acid and the hydroxyl on carbon-6 of glucose was measured. Active site residues are displayed as sticks. C. Surface view of the active site with salicylic acid docked to show the space of the active site cavity. D. Surface view of the active site with capsaicin docked to show the full extent of capsaicin binding both close to the site of glycosyl transfer and in the distal hydrophobic space within the pocket. Note that while the distances were drawn between the ligand and the hydroxyl on carbon-6 of glucose, the hydroxyl on carbon-1 that is covalently attached to the beta-phosphate of UDP is the hydroxyl on which the ligand is transferred.

In the virtual screen of the *A. thaliana* UGT74F2 with UDP-glucose bound, we found an overall trend in our inspection of each compound (Table 4). Notably, small, aromatic compounds, such as 2,4-dihydroxybenzoic acid and salicylic acid, have a high binding affinity (−6.5 and −6 kcal/mol, respectively) and a hydroxyl or carboxyl oxygen on these molecules is a short distance from a hydroxyl oxygen on the glucose molecule (approximately 2.8 Å and 2.7 Å, respectively). Aromatic amino acids in the active site (Tyr13, Phe70, Phe113, Tyr180, and Trp364) likely stabilize intramolecular interactions between substrates and the protein through π-π stacking

interactions (Fig. 2). His18 is a conserved catalytic residue that hydrogen bonds with substrates to position them for the glycosyl transfer reaction (George Thompson et al., 2017). When salicylic acid is the substrate, the protonation state of the basic active site amino acid His18 determines whether the carboxyl or hydroxyl are glucosylated by the enzyme. The carboxylate of SA likely forms hydrogen bonding interactions with the side-chain hydroxyl of Thr15, while the carboxylate of 2,4-dihydroxybenzoic acid is on the opposite side of the active site and is predicted to form hydrogen bonds with the side-chain hydroxyl of Thr365. These two threonine residues have been found to contribute to substrate binding conformation.

These results are consistent with previous functional data that found that UGT74F2 can glycosylate the carboxylate and the hydroxyl of SA (2-hydroxybenzoic acid), anthranilate, benzoic acid, and 3-hydroxybenzoic acid (George Thompson et al., 2017; Lim et al., 2002; Quiel & Bender, 2003). However, activity was not seen with 4-hydroxybenzoic acid, 2,4-dihydroxybenzoic acid or any other dihydroxybenzoates that were tested (Lim et al., 2002). This discrepancy highlights the importance of validating the results from a virtual screen using functional assays.

Interestingly, the molecule with the highest binding affinity that was within 5 Å of UDP-glucose was the alkaloid capsaicin (−7.2 kcal/mol). Capsaicin is an alkaloid produced in the *Capsicum* genus (Solanaceae) and is responsible for the pungent and spicy sensation associated with eating hot peppers.

Capsaicin bound the active site with a hydroxyl on the aromatic ring positioned 2.8 Å away from the hydroxyl on carbon-6 of glucose, and the hydrophobic 8-methyl-6-nonenoyl moiety oriented near hydrophobic residues in the active site pocket (Phe 43, Val 44, and Phe 70) (Fig. 2). Capsaicin glucosides have been detected in *Capsicum annuum* peppers, and one glucoside, capsaicin-*β*-D-glucopyranoside, was found to have 1/100th of the pungency of capsaicin (Higashiguchi, Nakamura, Hayashi, & Kometani, 2006; Kometani, Tanimoto, Nishimura, Kanbara, & Okada., S., 1993). While insect UGTs that glycosylate capsaicin as a detoxification mechanism have been characterized, a plant enzyme that glycosylated capsaicin remains to be identified (Ahn et al., 2011).

Other compounds in the table that have not been screened for in vitro activity but bind in the active site in the virtual screen include the trihydroxybenzoate gallic acid, the tryptophan catabolism intermediate 3-hydroxyanthranilate, the auxinic hormone phenylacetic acid, and the phenylpropanoid biosynthetic intermediate *p*-coumaryl alcohol. Additionally,

three herbicides were predicted to bind the active site of the enzyme: clopyralid (3,6-dichloro-2-pyridinecarboxylic acid), glyphosate (*N*-(phosphonomethyl)glycine), and metam (methylcarbamodithioic acid). The three herbicides have lower binding affinities than the known substrates, but because this is a computational screen, the binding of these molecules to UGT74F2 should be confirmed experimentally. It may be that enzyme activity is unaffected by these herbicides, or the herbicides could serve as competitive inhibitors by binding and blocking substrates. Assays would also be needed to determine whether the carboxylate of clopyralid, the carboxylate of glyphosate, or the thiol of metam can be glycosylated, which may serve as a mechanism for herbicide resistance if the glycosylated herbicide has reduced toxicity (Gaines et al., 2020).

The virtual screen that included UDP bound in the active site generated more compounds that were able to bind in the active site than in the UDP-glucose structure, which is likely because the glucose moiety occupies a large space in the active site and excludes larger molecules (Table 5). For the UDP-bound enzyme, a cutoff of 6.0 Å between hydroxyl on carbon-6 of glucose and an atom on the compound that could be glycosylated was selected (Fig. 1). Known substrates, including anthranilate, salicylic acid, and 3-hydroxybenzoic acid, had predicted binding affinities of −5.9, −5.6, and −5.5 kcal/mol, respectively (Table 5). Unsurprisingly the results of docking with UDP and UDP-glucose bound in the active site varied, which highlights the importance of testing multiple cosubstrates, coproducts, or cofactors.

5. Limitations of virtual screens

While docking and virtual screens are an excellent way to computationally predict how a small molecule will interact with an enzyme of interest, this technique is not without its limitations. One of the limitations of the virtual screening method provided is that the protein or enzyme of interest will be screened for ligand binding based on a single, rigid protein conformation. Regardless of the source of the protein structure (i.e., X-ray crystal structure, homology model, etc.), flexibility of the ligand will not be taken into account during the virtual screen. Virtual screen results may not accurately depict the molecular interactions that would occur between the protein and ligand in another conformation. If protein flexibility is a key consideration in generating reliable docking results for the protein of interest, this AutoDock Vina-based virtual screen could be followed with a

second screen using DOCK 6 with the AMBER scoring function, which performs molecular dynamics simulations (Allen et al., 2015; Maia, Assis, de Oliveira, da Silva, & Taranto, 2020). However, there are limitations to this approach as well, namely, the input preparation and the amount of time needed to run a more computationally intensive program.

Another limitation inherent to virtual screening is that there are false positives, and perhaps false negatives, within the data (Maia et al., 2020). For example, metabolites that are known to not exhibit activity with the *A. thaliana* UGT74F2, such as 3,4-dihydroxybenzoic acid, have a higher predicted binding affinity than known substrates, such as salicylic acid and anthranilate (Table 4). Like any computational screen, in vitro or in vivo experiments are needed to complement and confirm the results. For example, if the metabolite is a putative substrate, then the enzyme could be purified and assayed for activity with the metabolite(s) of interest. Gene knockout or knockdown experiments coupled with metabolomics could be used to determine the chemical phenotype in the absence of the enzyme. Depending on the number of ligands included in the screen, one may end up with a list of tens of metabolites to screen for activity, many of which may not serve as substrates. On the other hand, without knowing that anthranilate and salicylic acid are substrates for UGT74F2, one may have made an arbitrary cutoff for binding affinities that excluded these compounds from in vitro analyses, which would have meant that the substrates of the enzyme may not have been identified.

A key consideration for virtual screening is the space that cofactors and cosubstrates occupy in the active site. In the case presented here, docking was performed with both UDP-glucose, which is a cosubstrate, and the product UDP. While there were similarities between the results (Tables 4 and 5), there are also notable differences. The UGT74F2 active site was able to accommodate larger molecules when only UDP was bound, whereas smaller aromatic compounds predominantly bound the active site when UDP-glucose was present. Because of the speed of conducting a virtual screen, re-screening enzymes with multiple cofactors or cosubstrates is feasible.

Many simulation-based softwares are limited in their ability to model complex systems to perfect accuracy. Furthermore, the ability of protein-ligand docking programs such as AutoDock to provide accurate results is highly dependent upon the accuracy of the binding site of the structural model (Cross et al., 2009). Most protein modeling programs use homology-based techniques, where an unknown protein of interest is

modeled using a known protein with a solved crystal structure as a template. In some previous pharmacological studies, unknown proteins that shared >50% sequence identity with the template structure were deemed eligible for docking (Hillisch, Pineda, & Hilgenfeld, 2004). More recent findings suggest that the 50% homology cutoff is not an effective method for assessing docking eligibility, but with recent advances in predictive modeling programs like Alpha Fold, highly accurate protein structures have never been more accessible (Bordogna, Pandini, & Bonati, 2011; Jumper et al., 2021). As these protein structures continue to increase in accuracy, so do the capabilities of docking programs to accurately evaluate and model protein–ligand interactions.

Protein-ligand docking programs often suffer from an inability to predict the influence of water molecules on protein–ligand interactions (Verdonk et al., 2005). This is especially problematic for enzymes that contain water molecules in their active site, as these molecules may play a vital role in facilitating catalysis. During docking, optimization of water molecule orientation has been found to significantly increase docking accuracy (Roberts & Mancera, 2008). In addition to increasing docking accuracy, inclusion of water molecules can also be a useful tool for discovering substrates that are able to displace water molecules from the active site (Verdonk et al., 2005). Some docking programs, such as GOLD, can be modified to factor water mediation and displacement during protein-ligand docking. More recently, AutoDock Vina version 1.2.0 has been updated to support the modeling of explicit water molecules via the hydrated docking protocol (Eberhardt et al., 2021). This method produces a more accurate portrayal of water–protein–ligand interactions; however, it comes at a high computational cost compared to the default solvent models.

6. Conclusions

While virtual screening is commonly used in the pharmaceutical and biotechnology industries for drug and small-molecule inhibitor design, there is an untapped potential for this computational method for exploring the function of plant enzymes. Now with advances in protein structure prediction and the availability of ready-to-dock ligands, conducting a virtual screen can be free, easy, relatively fast and complementary to gene-expression-based methods that have been used extensively to identify candidate enzymes in metabolic pathways.

Before starting in vitro or in vivo experiments to confirm the function of a metabolic enzyme, conducting a virtual screen of the putative substrates and/or products of the enzymes of interest may provide an additional mechanism to narrow down lists of candidate genes. Additionally, virtual screens may be applied to enzymes with known functions to identify alternative substrates, identify natural inhibitors, or for designing inhibitors, like herbicides. However, like any computational tool, the results of virtual screens need to be confirmed experimentally.

Specialized metabolites have revolutionized the medical, cosmetic, and agricultural industries, however, enzyme promiscuity in specialized metabolic pathways also poses significant threats, such as the evolution of herbicide resistance (Abdollahi et al., 2021). While the evolution of enzyme specificity played a major role in increasing metabolic efficiency, enzyme promiscuity continues to drive the emergence of secondary metabolic pathways, which often provide a selective advantage to the host organism (Leong & Last, 2017). Understanding the mechanisms that drive enzyme promiscuity is key to navigating these challenges and may even allow for improved manipulation of valuable specialized metabolic pathways. Protein-based computational techniques such as virtual screening are excellent tools for probing individual metabolic enzymes for large pools of potential substrates. High-throughput analysis of structurally similar substrates will reveal key details about enzyme activity and may enable the development of novel pathways for natural product biosynthesis.

Acknowledgments

Research in the Holland lab is supported by the U.S. National Science Foundation (MCB-2214883) and by Williams College. The Perl script in the methods was made publicly available by Dr. Babajan Banaganapalli. The authors thank current and former lab members for testing the virtual screening protocols.

References

Abdollahi, F., Alebrahim, M. T., Ngov, C., Lallemand, E., Zheng, Y., Villette, C., et al. (2021). Innate promiscuity of the CYP706 family of P450 enzymes provides a suitable context for the evolution of dinitroaniline resistance in weed. *The New Phytologist*, *229*(6), 3253–3268.

Ahn, S. J., Badenes-Pérez, F. R., Reichelt, M., Svatoš, A., Schneider, B., Gershenzon, J., et al. (2011). Metabolic detoxification of capsaicin by UDP-glycosyltransferase in three Helicoverpa species. *Archives of Insect Biochemistry and Physiology*, *78*(2), 104–118.

Akere, A., Chen, S. H., Liu, X., Chen, Y., Dantu, S. C., Pandini, A., et al. (2020). Structure-based enzyme engineering improves donor-substrate recognition of Arabidopsis thaliana glycosyltransferases. *The Biochemical Journal*, *477*(15), 2791–2805.

Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., et al. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, *36*(15), 1132–1156.

Baldwin, I. T. (1989). Mechanism of damage-induced alkaloid production in wild tobacco. *Journal of Chemical Ecology*, *15*(5), 1661–1680.

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2014). The ChEMBL bioactivity database: An update. *Nucleic Acids Research*, *42*(D1), D1083–D1090.

Bordogna, A., Pandini, A., & Bonati, L. (2011). Predicting the accuracy of protein-ligand docking on homology models. *Journal of Computational Chemistry*, *32*(1), 81–98.

Caputi, L., Franke, J., Farrow, S. C., Chung, K., Payne, R., Nguyen, T. D., et al. (2018). Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science (New York, N.Y.)*, *360*(6394), 1235–1239.

Cross, J. B., Thompson, D. C., Rai, B. K., Baber, J. C., Fan, K. Y., Hu, Y., et al. (2009). Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling*, *49*(6), 1455–1474.

Delli-Ponti, R., Shivhare, D., & Mutwil, M. (2021). Using gene expression to study specialized metabolism—A practical guide. *Frontiers in Plant Science*, *11*, 2074.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, *61*(8), 3891–3898.

Fang, C., Fernie, A. R., & Luo, J. (2019). Exploring the diversity of plant metabolism. *Trends in Plant Science*, *24*(1), 83–98.

Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, *11*(5), 905–919.

Gaines, T. A., Duke, S. O., Morran, S., Rigon, C., Tranel, P. J., Küpper, A., et al. (2020). Mechanisms of evolved herbicide resistance. *The Journal of Biological Chemistry*, *295*(30), 10307–10330.

George Thompson, A. M., Iancu, C. V., Neet, K. E., Dean, J. V., & Choe, J. Y. (2017). Differences in salicylic acid glucose conjugations by UGT74F1 and UGT74F2 from Arabidopsis thaliana. *Scientific Reports*, *7*, 46629.

Higashiguchi, F., Nakamura, H., Hayashi, H., & Kometani, T. (2006). Purification and structure determination of glucosides of capsaicin and dihydrocapsaicin from various capsicum fruits. *Journal of Agricultural and Food Chemistry*, *54*(16), 5948–5953.

Hillisch, A., Pineda, L. F., & Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. *Drug Discovery Today*, *9*(15), 659–669.

Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., et al. (2020). ZINC20—A free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, *60*(12), 6065–6073.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kautsar, S. A., Suarez Duran, H. G., & Medema, M. H. (2018). Genomic identification and analysis of specialized metabolite biosynthetic gene clusters in plants using PlantiSMASH. *Methods in molecular biology (Clifton, N.J.)*, *1795*, 173–188.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845–858.

Kilgore, M. B., Holland, C. K., Jez, J. M., & Kutchan, T. M. (2016). Identification of a noroxomaritidine reductase with Amaryllidaceae alkaloid biosynthesis related activities. *Journal of Biological Chemistry*, *291*(32), 16740–16752.

Kim, J., & Buell, C. R. (2015). A revolution in plant metabolism: Genome-enabled pathway discovery. *Plant Physiology*, *169*(3), 1532–1539.
Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213.
Kometani, T., Tanimoto, H., Nishimura, T., Kanbara, I., & Okada, S. (1993). Glucosylation of capsaicin by cell suspension cultures of *Coffea arabica*. *Bioscience, Biotechnology, and Biochemistry*, *57*(12), 2192–2193.
Leong, B. J., & Last, R. L. (2017). Promiscuity, impersonation and accommodation: Evolution of plant specialized metabolism. *Current Opinion in Structural Biology*, *47*, 105–112.
Lim, E. K., Doucet, C. J., Li, Y., Elias, L., Worrall, D., Spencer, S. P., et al. (2002). The activity of Arabidopsis glycosyltransferases toward salicylic acid, 4-hydroxybenzoic acid, and other benzoates. *The Journal of Biological Chemistry*, *277*(1), 586–592.
Maeda, H. A., & Fernie, A. R. (2021). Evolutionary history of plant metabolism. *Annual Review of Plant Biology*, *72*, 185–216.
Maia, E., Assis, L. C., de Oliveira, T. A., da Silva, A. M., & Taranto, A. G. (2020). Structure-based virtual screening: From classical to artificial intelligence. *Frontiers in Chemistry*, *8*, 343.
Murata, K., Kitano, T., Yoshimoto, R., Takata, R., Ube, N., Ueno, K., et al. (2020). Natural variation in the expression and catalytic activity of a naringenin 7-O-methyltransferase influences antifungal defenses in diverse rice cultivars. *The Plant Journal*, *101*(5), 1103–1117.
Nützmann, H. W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters - from genetics to genomics. *The New Phytologist*, *211*(3), 771–789.
O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*(1), 1–14.
Polturak, G., Liu, Z., & Osbourn, A. (2022). New and emerging concepts in the evolution and function of plant biosynthetic gene clusters. *Current Opinion in Green and Sustainable Chemistry*, *33*, 100568.
Quiel, J. A., & Bender, J. (2003). Glucose conjugation of anthranilate by theArabidopsis UGT74F2 glucosyltransferase is required for tryptophan mutant blue fluorescence. *Journal of Biological Chemistry*, *278*(8), 6275–6281.
Rester, U. (2008). From virtuality to reality - virtual screening in lead discovery and lead optimization: A medicinal chemistry perspective. *Current Opinion in Drug Discovery & Development*, *11*(4), 559–568.
Rhee, S. Y., & Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, *19*(4), 212–221.
Roberts, B. C., & Mancera, R. L. (2008). Ligand-protein docking with water molecules. *Journal of Chemical Information and Modeling*, *48*(2), 397–408.
Saito, K., & Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annual Review of Plant Biology*, *61*, 463–489.
Schultz, B. J., Kim, S. Y., Lau, W., & Sattely, E. S. (2019). Total biosynthesis for milligram-scale production of etoposide intermediates in a plant chassis. *Journal of the American Chemical Society*, *141*(49), 19231–19235.
Sherp, A. M., Westfall, C. S., Alvarez, S., & Jez, J. M. (2018). Arabidopsis thaliana GH3.15 acyl acid amido synthetase has a highly specific substrate preference for the auxin precursor indole-3-butyric acid. *Journal of Biological Chemistry*, *293*(12), 4277–4288.
Singh, A., Menéndez-Perdomo, I. M., & Facchini, P. J. (2019). Benzylisoquinoline alkaloid biosynthesis in opium poppy: An update. *Phytochemistry Reviews*, *18*(6), 1457–1482.

Smith, S. D., Angelovici, R., Heyduk, K., Maeda, H. A., Moghe, G. D., Pires, J. C., et al. (2019). The renaissance of comparative biochemistry. *American Journal of Botany*, *106*(1), 3–13.

Sterling, T., & Irwin, J. J. (2015). ZINC 15- -ligand discovery for everyone. *Journal of Chemical Information and Modeling*, *55*(11), 2324–2337.

Sun, Y., Shang, L., Zhu, Q. H., Fan, L., & Guo, L. (2022). Twenty years of plant genome sequencing: Achievements and challenges. *Trends in Plant Science*, *27*(4), 391–401.

Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, *31*(2), 455–461.

Verdonk, M. L., Chessari, G., Cole, J. C., Hartshorn, M. J., Murray, C. W., Nissink, J. W., et al. (2005). Modeling water molecules in protein-ligand docking using GOLD. *Journal of Medicinal Chemistry*, *48*(20), 6504–6515.

Villoutreix, B. O., Renault, N., Lagorce, D., Sperandio, O., Montes, M., & Miteva, M. A. (2007). Free resources to assist structure-based virtual ligand screening experiments. *Current Protein & Peptide Science*, *8*(4), 381–411.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1), W296–W303.

Wisecaver, J. H., Borowsky, A. T., Tzin, V., Jander, G., Kliebenstein, D. J., & Rokas, A. (2017). A global Coexpression network approach for connecting genes to specialized metabolic pathways in plants. *The Plant Cell*, *29*(5), 944–959.