THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Letter

Proteome-Informed Machine Learning Studies of Cocaine Addiction

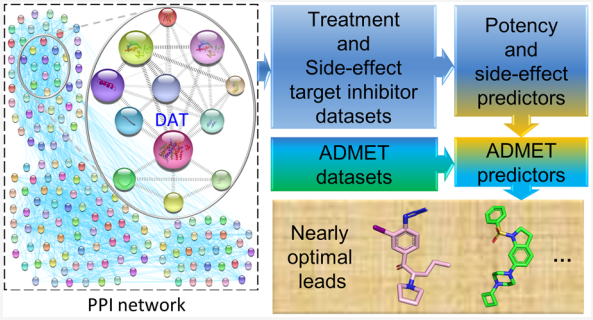
Kaifu Gao, Dong Chen, Alfred J. Robison, and Guo-Wei Wei*

Cite This: J. Phys. Chem. Lett. 2021, 12, 11122−11134

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: No anti-cocaine addiction drugs have been approved by the Food and Drug Administration despite decades of effort. The main challenge is the intricate molecular mechanisms of cocaine addiction, involving synergistic interactions among proteins upstream and downstream of the dopamine transporter. However, it is difficult to study so many proteins with traditional experiments, highlighting the need for innovative strategies in the field. We propose a proteome-informed machine learning (ML) platform for discovering nearly optimal anti-cocaine addiction lead compounds. We analyze proteomic protein−protein interaction networks for cocaine dependence to identify 141 involved drug targets and build 32 ML models for cross-target analysis of more than 60,000 drug candidates or experimental drugs for side effects and repurposing potentials. We further predict their ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties. Our platform reveals that essentially all of the existing drug candidates fail in our cross-target and ADMET screenings but identifies several nearly optimal leads for further optimization.



Substance use disorders (SUD) involving alcohol, opioids, cocaine, etc., adversely affect a growing population of individuals and families worldwide, constituting a significant socioeconomic burden with increasing medical expenses and crime. Psychostimulants, especially cocaine, account for a large portion of SUD and impact millions of lives. In the United States alone, among the 70,630 SUD-related deaths in 2019, 15,883 were due to cocaine addiction. The hazard from cocaine addiction and subsequent mortality calls for effective medications. However, currently no cocaine addiction medications have been approved by the U.S. Food and Drug Administration (FDA).[1]

The psychotropic properties of cocaine primarily derive from blocking the dopamine transporter (DAT). Specifically, cocaine blocks DAT and prevents dopamine reuptake from the synaptic cleft into the presynaptic axon terminal. As a result, a higher dopamine level in the synaptic cleft promotes the activation of dopamine receptors in the postsynaptic neuron, which generates euphoria and arousal.[2] Among these dopamine receptors, the D_3 dopamine receptor (D_3R) plays a critical role in the reward and addiction of cocaine because the population of D_3R in the mesolimbic reward system is large.[3] Therefore, D_3R may be an important target for treating cocaine addiction. Among other dopamine receptors, D_1R and D_2R are the most abundant in the brain. D_1R along with other D_1-like receptors stimulates intracellular cyclic adenosine monophosphate (cAMP) levels.[4] The functions of D_1-like receptors are to regulate the growth of neurons, some D_2R-mediated events, and other behaviors.[5] D_2R, D_3R, and D_4R belong to the group of D_2-like receptors and inhibit intracellular cAMP levels. D_2R intimately joins in the circuitry of motor control, and it is the main target of most antipsychotic drugs.[6] D_4R relates to many neurological and psychiatric conditions,[7] including schizophrenia and bipolar disorder, attention deficit hyperactivity disorder (ADHD), addictive behaviors, Parkinson's disease, and eating disorders such as anorexia nervosa. Some studies also suggest that D_1R, D_2R, and D_4R are involved in the changes of locomotor activity induced by cocaine and other psychostimulants.[8]

Cocaine also blocks the serotonin transporter and norepinephrine transporter, inhibiting the reuptake of serotonin and norepinephrine and thus increasing the level of the activation of serotonin and norepinephrine receptors. Additionally, cocaine exposure could regulate opioid receptors and endogenous opioid peptides[9] and may also affect the selection of G-protein versus β-arrestin pathways.[10,11] Repeated use of psychostimulants alters gene expression throughout the brain, including in the nucleus accumbens, a critical center for reward processing. Frequent cocaine exposure increases the level of expression of the transcription factor ΔFosB and brain-derived neurotrophic factor (BDNF), which in turn regulate gene expression to alter both dendritic and synaptic morphology and function in the nucleus

Received: September 23, 2021
Accepted: November 5, 2021
Published: November 9, 2021

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

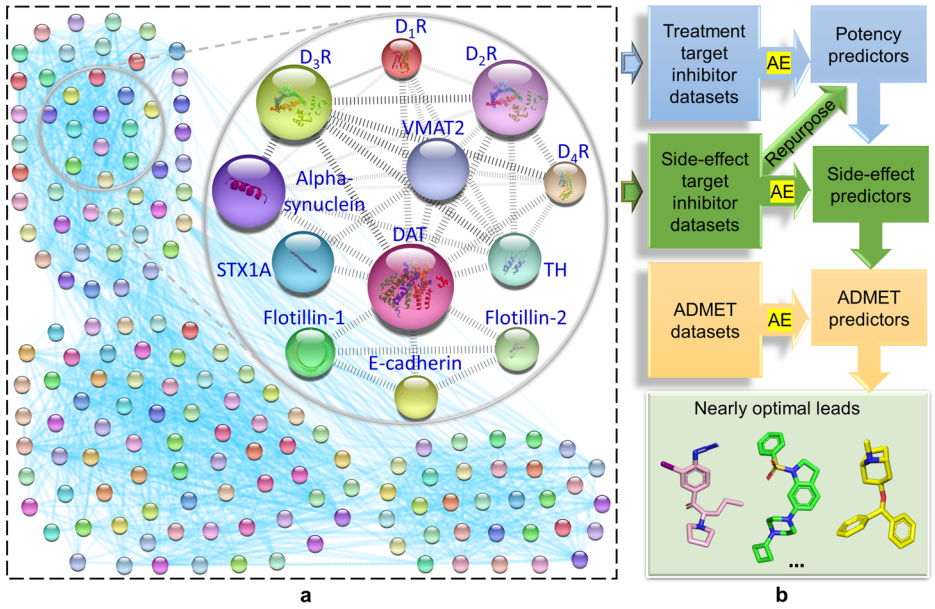


Figure 1. DAT-centered global and core PPI networks and proteome-informed ML workflow for anti-cocaine addiction drug discovery. (a) DAT-centered global and core PPI networks. (b) Our proteome-informed ML workflow for anti-cocaine addiction drug discovery. First, the data sets inferred by the PPI networks are collected, represented by a latent-vector fingerprint via an autoencoder, and used to construct binding affinity predictors. Second, the hits obtained from potency predictors, including those repurposing hits, are screened for potential side effects. Finally, the resulting promising candidates are further evaluated for ADMET properties to discover nearly optimal anti-cocaine addiction leads. Abbreviations: DAT, dopamine transporter; D_1R, dopamine receptor D1; D_2R, dopamine receptor D2; D_3R, dopamine receptor D3; D_4R, dopamine receptor D4; STX1A, syntaxin-1A; TH, tyrosine hydroxylase; VMAT2, vesicular monoamine transporter 2.

accumbens and prefrontal cortex,[12] likely driving the long-term compulsion for drug seeking and taking that underlies addiction.[13]

Currently, experimental medications against cocaine addiction mainly target DAT and D_3R.[1] (1) Atypical DAT inhibitors are studied widely. While cocaine and its analogues (typical DAT inhibitors) bind and stabilize outward-facing conformations of DAT,[14] atypical DAT inhibitors stabilize inward-facing conformations of DAT upon their binding. DAT with an inward-facing conformation is much harder for cocaine to bind (an approximate 100-fold loss of the potency of cocaine for the inward-facing conformation compared with that of the outward-facing conformation).[14,15] In other words, even binding affinities (BAs) for DAT are weaker than that of cocaine, and the pretreatment by atypical inhibitors can still prevent DAT from being blocked by cocaine. More importantly, atypical DAT inhibitors do not induce cocaine-like behaviors or addiction.[15] (2) Another promising approach against cocaine addiction involves D_3R antagonists and/or partial agonists. D_3R antagonists could effectively attenuate the motivation to earn psychostimulants and reduce relapse-related behaviors. D_3R partial agonists not only can functionally block the effect of cocaine addiction but also can elicit the partial activation of their receptor targets under abstinence conditions and thus potentially mitigate withdrawal effects.[16]

In addition to potency, the safety of cocaine addiction treatments must be carefully evaluated. One dangerous side-effect target for drug addiction treatments is the human *ether-a-go-go* (hERG) potassium channel, which could incur adverse side effects and even death. hERG generates the delayed rectifying potassium current. When a compound inhibits the hERG channel, it interferes with potassium current, prolongs the QT interval, and results in torsades de pointes (TdP), a potentially lethal ventricular tachycardia.[17] Thus, hERG poses a serious challenge to drug development because it can easily attract small compounds, especially those with protonatable amines and aromatic groups, a hallmark of many neurotransmitter transport inhibitors and GPCR ligands.[18] The hERG blockade was a popular reason for drug withdrawals in the 1990s and early 2000s. Therefore, in early 2000, the FDA included the hERG side effect in their updated regulations: the TdP liability of drug candidates must be evaluated *in vivo* or *in vitro* in phase 1 clinical trials.[19]

The mechanism of cocaine addiction is very complicated, involving far more targets than DAT, D_3R, and hERG. All of the proteins upstream and downstream of DAT functions could be impacted by cocaine, which covers a large number of proteins and interactions as shown in Figure 1a. On one hand, these proteins can become potential treatment targets for cocaine addiction. On the other hand, blocking these proteins also probably brings cocaine-like symptoms or other severe off-target effects. Therefore, these proteins could be critical sources of side effects. Thus, we need to systematically investigate potential compounds that inhibit different cocaine addiction targets, as well as the putative side effects from agents blocking these targets.

One method for systematically unveiling potential treatment and critical side-effect targets is to examine sizable protein−protein interaction (PPI) networks on the proteome scale. A PPI network accounts for not only direct (physical and chemical) interactions but also indirect (functional) association,[20] in which a connection represents two proteins jointly contributing to a specific biological function even without direct physical or chemical interaction. As a result, a proteomic PPI network is a suitable tool for systematically searching a large number of proteins relating to a specific disease, providing a "pool" of potential treatments and critical side-effect targets, such as cocaine addiction in this work. The

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

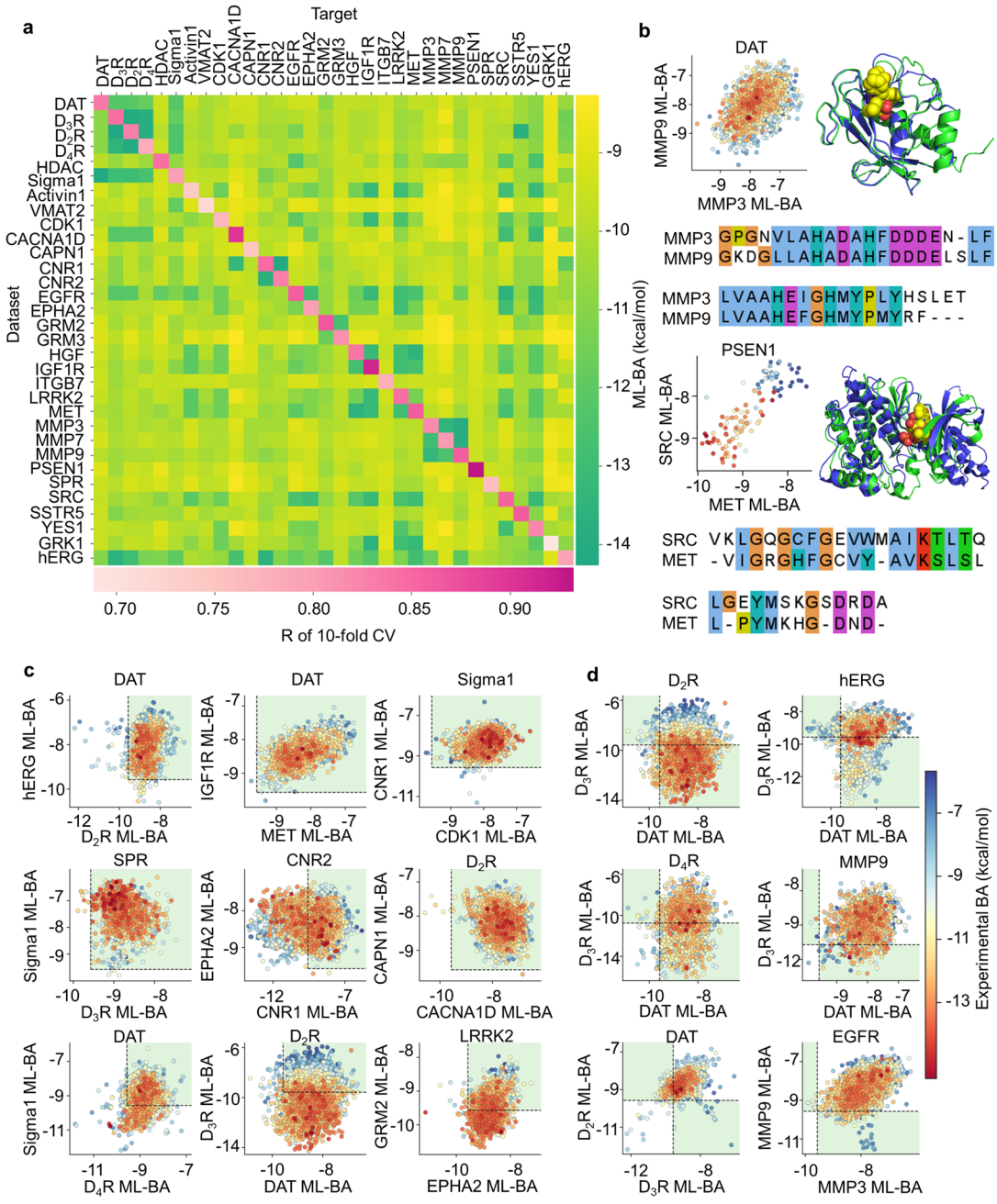


Figure 2. Cross-target BA prediction systematically suggesting side effects and repurposing potential. (a) Heat map of cross-target BA prediction indicating the inhibitor specificity on each data set. In each row, the diagonal element shows the Pearson correlation coefficients of 10-fold cross-validation (*R* of 10-fold CV) on the machine learning-predicted BAs (ML-BAs) of each data set. Other elements represent the highest ML-BAs among the inhibitors in each data set to other targets. (b) Two examples of positive cross-target BA correlation revealing binding-site similarity of MMP3 with MMP9 and MET with SPC. In each example, the ML-BA correlation plot of one data set to the target pair, the three-dimensional alignment of the two target proteins, and the two-dimensional sequence alignment of their binding sites are given. The Protein Data Bank entries of the protein structures are 1B3D, 1GKC, 1R0P, and 1KSW for MMP3, MMP9, MET, and SPC, respectively. In each ML-BA correlation plot, the title is the name of the data set. The colors of points represent the experimental BA for the designated target. The *x*- and *y*-axes indicate the ML-BA for two other proteins. (c) Nine typical examples of cross-target BA predictions of potential side effects. The first, second, and third rows exhibit the examples with substantial side effects of potent inhibitors on zero, one, and two targets, respectively. The green frames outline the optimal ranges without side effects on both targets ($x > -9.57$ kcal/mol, and $y > -9.57$ kcal/mol). (d) Six typical examples of cross-target BA prediction suggesting repurposing potential. In these examples, some weak inhibitors of their designated targets are predicted to have high BAs (i.e., low BA values) to other proteins. The two green frames in each subplot outline the BA domains with repurposing potential, which represent compounds that have the potency to one target (BA values of <−9.57 kcal/mol) and do not show strong side effects on the other target (BA values of >−9.57 kcal/mol).

String v11 database collects a large number of protein−protein interactions involving a total of 24,584,628 proteins from 5090 organisms..[20] One can extract the large-scale PPI network for a specific human protein from String.

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

Traditional *in vivo* or *in vitro* experiments are too time-consuming and expensive to test all of the proteins in a proteomic PPI network efficiently. Additionally, large-scale experiments on animals could raise important ethical concerns. For large-scale assays, machine learning/deep learning (ML/DL) technologies are promising, at least for initial evaluation and screening. ML/DL technologies have been widely applied at different stages of drug design and discovery.[21] ML/DL could help to predict drug potency,[21] repurpose existing drugs to new diseases,[22] and even generate new druglike compounds for further screening.[23] ML/DL methods have also been established for the lead optimization of various druggable properties,[24,25] including solubility, partition coefficient, toxicity, pharmacokinetics, and pharmacodynamics. These technologies could largely reduce the need for time-consuming and expensive experiments and thus accelerate drug discovery, significantly benefiting human health and welfare.

In this work, we designed a proteome-informed ML/DL workflow to discover nearly optimal anti-cocaine addiction leads, as shown in Figure 1. First, we extracted a proteomic PPI network of cocaine addiction from the String database[20] to infer 141 potential treatment and critical side-effect targets. Although more side-effect targets outside the PPI network should be considered, we limit our effort to critical ones revealed by the PPI network in this work. Second, for targets in the network, the associated molecules are presented by latent-vector fingerprints (LV-FPs) via an autoencoder (AE) and built ML/DL-based BA predictors. Third, we carry out cross-target BA predictions of more than 60,000 associated compounds to screen possible side effects and repurposing potentials. Interestingly, the correlation between predicted BAs for different targets could reveal binding-site similarities among targets, which is a byproduct of our proteome-informed ML/DL workflow. Finally, we applied ML-based models to further evaluate the pharmacokinetic properties, i.e., absorption, distribution, metabolism, excretion, and toxicity (ADMET) as well as synthesizability. These evaluations, together with the potency and side-effect analysis, form a series of filters to screen nearly optimal lead compounds. Finally, we also study atypical or typical inhibition of these nearly optimal leads via induced-fit docking.

DAT is well-known as the critical direct target of cocaine. To study the PPI network of cocaine addiction, we input "DAT" into the String database and extracted a global network and a core network of DAT interactions (see Figure 1a). The global network contains 141 nodes and 1696 edges. The core network considers only the proteins having direct known interactions with DAT, which leads to 12 nodes and 29 edges. The global network could be decomposed into three clusters, implying these proteins involve three different primary functions. The core network resides in one cluster, with 12 critical proteins in the biochemical pathways of cocaine addiction.

Apart from DAT, VMAT2 is another critical node in the core network. VMAT2 is a transport protein integrated into the membrane of synaptic vesicles of presynaptic neurons. Its main function is to transport monoamines, especially neurotransmitters such as dopamine, norepinephrine, serotonin, and histamine, from the cytosol into synaptic vesicles, which then release the neurotransmitters into synapses as chemical messages to postsynaptic neurons. Many psychostimulants such as cocaine interact with VMAT2, which emphasizes its clinical significance.[26] Moreover, α-synuclein is a neuronal protein that plays several roles in synaptic activity, such as regulation of synaptic vesicle trafficking and subsequent neurotransmitter release. It participates as a monomer in synaptic vesicle exocytosis by enhancing vesicle priming, fusion, and dilation of exocytotic fusion pores. Cocaine abusers typically have overexpression of α-synuclein in dopamine neurons.[27] TH is the enzyme responsible for catalyzing the conversion of the amino acid L-tyrosine to L-3,4-dihydroxyphenylalanine (L-DOPA), which is a precursor for dopamine. Studies also suggested cocaine administration could increase TH enzyme activity.[28] STX1A is a nervous system-specific protein implicated in the docking of synaptic vesicles with the presynaptic plasma membrane. E-Cadherin is a type of cell adhesion molecule that is important in the formation of adherens junctions to bind cells to each other. Flotillin-1 and -2 are ubiquitously expressed, evolutionarily conserved peripherally membrane-associated proteins. Flotillins are found to regulate the membrane mobility of DAT.

The 12 aforementioned proteins constitute the core network of cocaine addiction. Their mutual interactions in the network also indicate DAT is the most important node, which connects the upper and lower parts of the network. Network analysis shows that DAT has the highest degree (11) among all of the nodes and is a hub of the core network. Additionally, the closeness centrality of DAT is as high as 1.000, which also suggests its full connection to all of the other proteins in the core network. More importantly, the betweenness centrality of DAT (0.470) is higher than those of any other nodes, suggesting other than the hub, DAT is also a critical bottleneck. In other words, DAT is a bridge of the network, and almost half of the interactions must be via DAT. If DAT is removed, the communication between the upper and lower parts of the core network will be essentially cut off.

The connections also reveal the importance of VMAT2 and α-synuclein. Their degrees are both 8, and the closeness centrality values are both 0.786. They have connections to all nodes in the upper part of the network, and both have betweenness centrality values of 0.033, forming shortcuts between other proteins. Three other proteins with positive betweenness centrality are D_3R, D_2R, and TH. Their betweenness centrality values are all 0.003, suggesting they play some roles as bottlenecks. For example, the shortest pathway between D_1R and D_4R is through D_3R or D_2R. Their degree and closeness centrality are 7 and 0.733, respectively.

As mentioned previously, the global PPI network of cocaine addiction involves as many as 141 proteins, which not only play roles in cocaine addiction but also participate in other biological activities. A drug must be specific to its own target and not affect other protein functions to avoid side effects. In this section, through ML/DL models, we systematically predict inhibitor BAs to analyze side effects and repurposing potential.

We collect inhibitor data from the ChEMBL database.[29] We build ML models for 32 proteins that have sufficient inhibitor data to build such models. These models are used for drug repurposing and side-effect studies.

Figure 2a depicts the heat map of cross-target BA predictions for 32 targets. Each diagonal element shows the Pearson correlation coefficient (R) of 10-fold cross-validation (CV) of the ML BA predictions (ML-BAs) for the corresponding protein inhibitor data set. Three of 32 models have R values of >0.90, showing excellent accuracy. The R values of 21 ML models are >0.80. For example, the R value of the ML model for the DAT data set of 2877 compounds is

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

0.84. Only one model's R value is <0.70 (i.e., R = 0.69 for the GRK5 model). Therefore, these ML models are quite reliable.

In Figure 2a, elements right to the diagonal in each row are the maximum BA values of the data set of the diagonal element predicted by the corresponding models. For example, element (1,2) is the maximum BA value of the DAT data set (2877 compounds) predicted by the D_3R ML model. Elements below the diagonal in each column are the maximum BA values of other data sets predicted by the diagonal model. For example, element (2,1) is the maximum BA value of the D_3R data set (4685 compounds) predicted by the DAT ML model.

For a given drug candidate, its predicted high cross-target ML-BAs might suggest strong side effects. Among 992 cross-target screenings in Figure 2a, there are 330 potential side effects judged by a threshold BA of −9.57 kcal/mol (K_i = 0.1 μM, which is a broadly accepted threshold for high affinity[30]). Some side effects are due to highly similar targets, such as receptors D_3R, D_2R, and D_4R, cannabinoid receptors CNR1 and CNR2, glutamate metabotropic receptors GRM2 and GRM3, and matrix metallopeptidases MMP3, MMP7, and MMP9. Their high degress of sequence and structure similarity contribute their mutual side effects. However, mutual side effects are also found between seemingly unrelated proteins, such as DAT and Sigma1, etc.

We next investigate the positive correlations between predicted ML-BAs. Figure 2b exhibits two examples of cross-target BA correlations. The first example depicts compounds in the DAT data set binding to targets MMP3 and MMP9, which play an important role in cocaine relapse.[31] The correlation plot reveals an R value of 0.48 between their predicted BAs. The second and third plots are the three-dimensional (3D) alignment of the proteins and two-dimensional (2D) sequence alignment of their binding sites, respectively, which suggest these proteins, and especially their binding sites, are highly similar, with a binding-site sequence identity as high as 64.9%.

In addition to the targets from the same protein family leading to correlated BAs, we also found some seemingly unrelated proteins with correlated BAs, indicating their binding sites are similar as shown in Figure 2b. Although tyrosine-protein kinase met (MET) and proto-oncogene tyrosine-protein kinase src (SRC) are not of the same family, they are both kinases. MET is a tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to a hepatocyte growth factor ligand. SRC is a tyrosine kinase that is activated following the engagement of many different classes of cellular receptors. The alignment plots in Figure 2b reveals the binding domains of MET and SRC having conserved 3D conformations with a 2D sequence identity of 50.1%. Thus, the R values between the BAs of MET and SRC data sets are as high as 0.82. More examples can be found in Figure S1.

One significant application of cross-target BA predictions is to evaluate side effects and repurposing potentials. Our basic idea is to systematically predict the BAs of the inhibitors of one target by using the ML models of data sets for other proteins. It is desirable for a drug candidate to be highly specific, i.e., having a high BA for its target, and have weak side effects, i.e., having very low BAs for all other human proteins. Moreover, if a drug candidate interacts weakly with its designated target but is predicted to be potent at another unintended protein, then it has repurposing potential. Here, we carefully studied the 330 data set-target pairs with potential mutual side effects in Figure 2a. Panels c and d of Figure 2 depict some typical examples of our side-effect and repurposing detection through cross-target BA predictions. In each chart, three targets are involved: its designated target and two other potential side-effect targets.

Figure 2c exemplifies side-effect predictions. The first row illustrates active inhibitors having no serious side effects on either of two other targets. In the three plots, all active compounds, which are represented by red or even deep red points, are predicted to have low BAs (i.e., BA values of >−9.57 kcal/mol) for two other proteins. Therefore, we anticipated that these active inhibitors would not have strong side effects on two other targets studied. The second row contains examples with predicted side effects on one of two targets. For instance, the second plot in this row shows that the potent inhibitors of protein CNR2 are unlikely to bind to EPHA2. However, some of these potent inhibitors have strong predicted BAs (≈−12 kcal/mol) for CNR1. This potential side effect is expected, as CNR1 and CNR2 are similar cannabinoid receptors. The third row is the worst case in which side effects are predicted for both of the two other proteins. The most obvious cases are due to the kinship of the involved proteins, which are included in Figure S2. However, we also noticed that some inhibitors can still cause simultaneous side effects on unrelated targets, such as Sigma1 and D_4R in the first chart in this row.

In addition to side-effect evaluation, our cross-target BA predictions could also suggest repurposing potential as shown in Figure 2d. In each subplot of Figure 2d, some inactive inhibitors are predicted to be potent inhibitors of other proteins. For instance, in the first chart, some inactive D_2R inhibitors have high ML-BAs (BA values of <−9.57 kcal/mol) on DAT, suggesting these D_2R inhibitors are potential DAT inhibitors for further studies. In the second chart, some compounds inactive to hERG are predicted to be very potent to DAT, while some other inactive compounds are predicted to strongly block D_3R. In the fifth chart, some inactive DAT inhibitors are predicted to be potent inhibitors of D_3R or D_2R. More side-effect and repurposing examples are given in Figure S2.

Because DAT is the main target of cocaine and hERG is a critical side-effect target especially for neurotransmitter transport inhibitors and GPCR ligands, in this section, we focused on predicting the BAs of inhibitors from all of the other 30 data sets on DAT and hERG, which could allow evaluation of their hERG side effects and repurposing potential against DAT.

Avoiding hERG side effects is a priority for all drugs. Herein, we designate a more strict threshold of −8.18 kcal/mol (K_i = 1 μM) for any hERG side effect. As shown in Figure S3, most of the inhibitors in many data sets have predicted hERG BA values of over −8.18 kcal/mol, which suggests no serious hERG side effect. Especially for data sets GRM3, LRRK2, and SPR, almost all of the compounds in these data sets weakly bind to hERG with BA values of over −8.18 kcal/mol. However, a large number molecules in data sets D_2R, D_3R, and D_4R were predicted to have serious hERG side effects.

In the examination of the repurposing potential for DAT compounds, inactive inhibitors in data sets D_2R, D_3R, D_4R, EGFR, MET, and Sigma1 with experimental BA values of more than −9.57 kcal/mol are predicted to be potent at DAT with BA values of less than −9.57 kcal/mol for DAT. Some of these compounds possess a low potential for hERG side effects as shown in Figure S3. Of particular note is the Sigma1 data set containing 44 inactive inhibitors but predicted to strongly inhibit DAT with ML-BA values of less than −9.57 kcal/mol.

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

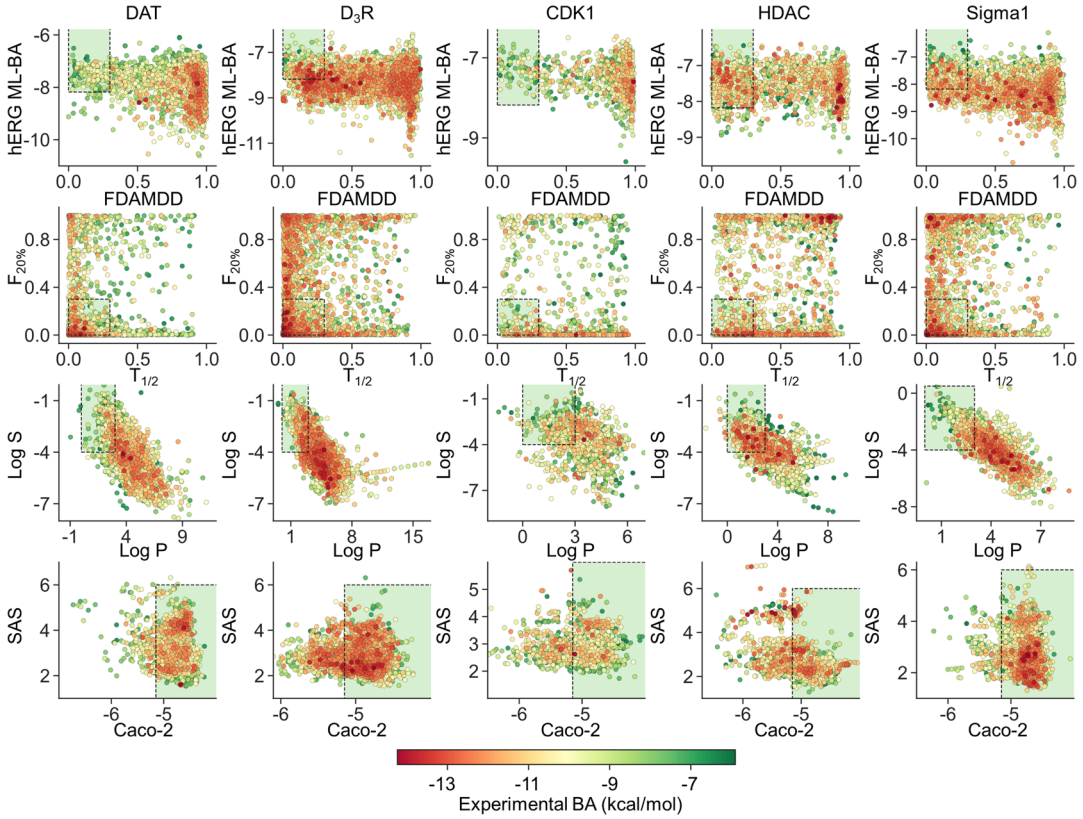


Figure 3. Druggable property screening based on ADMET properties, synthesizability, and hERG side effects on compounds from five critical protein data sets: DAT, D_3R, CDK1, HDAC, and Sigma1. The colors of the points represent the experimental BAs for these targets. The *x*- and *y*-axes show predicted ADMET properties, synthesizability, or hERG side effects. Green frames outline the optimal ranges of these properties and side effects.

Additionally, 11 of the 44 compounds were predicted to have weak hERG side effects with BA values of larger than −8.18 kcal/mol. Therefore, according to our predictions, these 11 compounds could potentially be repurposed for inhibiting DAT without strong hERG side effects. However, to qualify as nearly optimal lead compounds, further screenings for other side effects, ADMET properties, and synthesizability are indispensable.

Here we performed systematic screenings on ADMET properties, synthesizability, and hERG side effects. Figure 3 illustrates the example screening implemented on the data sets of five proteins (DAT, D_3R, CDK1, HDAC, and Sigma1) that play essential roles in cocaine addiction. The optimal ranges of ADMET properties and synthesizability are listed in Table 1, while a BA value of larger than −8.18 kcal/mol is applied as the required range for exempting hERG side effects.

Table 1. Optimal Ranges of the ADMET Properties and Synthesizability Considered in This Work

property	optimal range
FDAMDD	excellent, 0−0.3; medium, 0.3−0.7; poor, 0.7−1.0
$F_{20\%}$	excellent, 0−0.3; medium, 0.3−0.7; poor, 0.7−1.0
log *P*	proper range, 0−3 log mol/L
log *S*	proper range, −4− 0.5 log mol/L
$T_{1/2}$	excellent, 0−0.3; medium, 0.3−0.7; poor, 0.7−1.0
Caco-2	proper range, >−5.15
SAS	proper range, <6

Two critical properties for potential drug candidates are the FDA maximum recommended daily dose (FDAMDDs) and the BA for hERG (hERG_BA), representing the potential for toxicity and hERG side effects, respectively. The first row of Figure 3 depicts the distributions of these two properties of inhibitors from the five critical data sets. The green frames are the optimal domains of the two properties mentioned above. The colors of points represent experimental BA values for targets. According to this screening, all five data sets contain sufficient compounds with optimal toxicity and hERG side effects. However, for the CDK1 and DAT data sets, the optimal domains of toxicity and hERG side effects contain only very few potent inhibitors. This suggests ADMET properties and side effects must be considered before a new compound is synthesized.

The second row of Figure 3 illustrates the screening based on important absorption properties $T_{1/2}$ (half-life) and $F_{20\%}$ (bioavailability 20%). All five plots in the second row reveal that the optimal domain of $T_{1/2}$ and $F_{20\%}$ is only a small fraction of chemical space. However, for all five data sets, the small optimal domain does indeed contain some potent inhibitors.

The third row of Figure 3 displays the log *P* and log *S* screening. Log *P* and log *S* relate to the distribution of chemicals in human bodies. For all five targets, only a small portion of potent inhibitors can be found in the optimal domain, suggesting a huge waste of resources in early studies. Notably, there is an obvious line in the second subplot of this row, which is very unusual under natural conditions. This

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

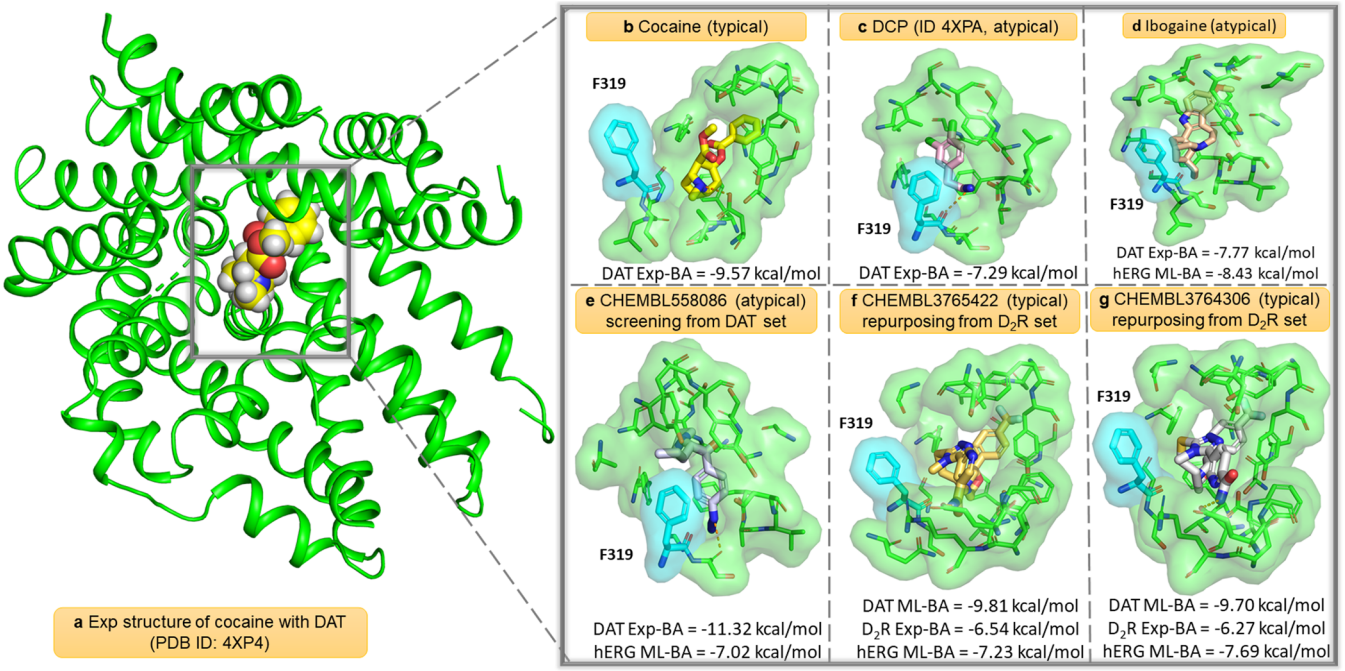


Figure 4. Experimental or docking structures of cocaine, DCP, ibogaine, and our nearly optimal lead compounds bound to *Drosophila* DAT revealing the molecular mechanism of typical or atypical inhibition. (a) Illustration of DAT inhibited by cocaine. (b and c) Modes of binding of cocaine and DCP, respectively, to *Drosophila* DAT from the experimental structures (Protein Data Bank entries 4XP4 and 4XPA,[52] respectively) revealing the molecular mechanism of typical or atypical inhibition. Because the orientation of Phe319 in *Drosophila* DAT (corresponding to Phe320 in human DAT) is the key to determining typical or atypical inhibition, Phe319 is enlarged and also colored blue. The comparison between panels b and c suggests, for typical inhibition in panel b, the side chain of Phe319 is stuck in the open state with the S1 pocket always open so that the occluded and inward-facing states of DAT cannot be formed. In contrast, Phe319 in panel c rotates to an occluded state, leading to atypical inhibition. (d) Docking pose of existing experimental medication ibogaine with DAT confirming its atypical inhibition. (e–g) Docking poses of our nearly optimal lead compounds suggesting their atypical or typical inhibition.

obvious line is probably caused by the intended optimization to improve log *P*.

The last row of Figure 3 depicts the Caco-2 and SAS screening. Caco-2 represents the cell permeability of compounds, while SAS describes how hard a compound is to synthesize. These plots indicate almost all of the compounds from the five data sets are not hard to synthesize, and approximately half of the compounds have good cell permeability. More importantly, many potent inhibitors are also in the optimal domain.

According to the literature,[1] some experimental medications are being investigated to treat cocaine addiction. Here we used our proteome-informed ML models to predict the side effects of these experimental medications. Ibogaine is a naturally occurring psychoactive substance that may have antiaddiction properties, and its docking structure and some experimental and predicted BAs are shown in Figure 4d. All of the 2D structures of existing experimental medications discussed in this work and their experimental or predicted BA values are available in Figures S4 and S5.

A typical DAT inhibitor such as cocaine binds and stabilizes an outward-facing conformation of DAT. However, an atypical DAT inhibitor stabilizes an inward-facing conformation of DAT upon its binding, which makes the binding of cocaine difficult.[15] Therefore, even with a weaker BA on DAT than cocaine, the pretreatment by an atypical inhibitor could still prevent DAT from being blocked by cocaine. More importantly, compared to typical DAT inhibitors such as cocaine, atypical DAT inhibitors increase the accessibility of residues in the cytoplasmic substrate permeation pathway and do not induce a cocaine-like behavior or addiction.[15]

Ibogaine is a natural psychoactive substance extracted from the plants in the Apocynaceae family such as *Tabernanthe iboga*, *Voacanga africana*, and *Tabernaemontana undulata* (see Figure 4d). Ibogaine was originally used in African spiritual ceremonies. However, its antiaddictive properties were accidentally discovered in 1962. Since then, ibogaine has been tested to treat SUD, especially for cocaine addiction. Now it is already approved for clinical use in The Netherlands, Canada, and Mexico.

Ibogaine inhibits DAT and SERT with IC_{50} values of 4.0 and 0.59 μM, respectively.[32] More importantly, ibogaine is an atypical inhibitor of DAT and SERT and has potential for treating cocaine addiction. However, its severe side effects and related deaths are of serious concern. Between 1990 and 2008, a total of 19 fatalities associated with the ingestion of ibogaine were reported, and six of these fatalities were caused by acute heart failure or cardiopulmonary arrest.[33] Our model predicted a moderate BA of ibogaine to hERG with a BA value of −8.43 kcal/mol, which suggests a risk for ibogaine to incur heart issues. Additionally, with predicted BA values of −9.71, −9.54, −9.46, and −9.43 kcal/mol with proteins YES1, LRRK2, GRM2, and FER, respectively, our models anticipated high risks of side effects on these proteins, especially for YES1 and FER, associated with severe diseases sarcoma and acute myeloid leukemia, respectively.

The structure of docking of ibogaine to DAT in Figure 4d reveals that, just like the crystal structure of cocaine with DAT in Figure 4b, ibogaine binds to DAT mainly through

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

hydrophobic interactions. Strong hydrogen interactions are absent.

Modafinil is a functional stimulant targeting DAT and NET,[34] which has been approved for the treatment of excessive sleepiness, such as narcolepsy or idiopathic hypersomnia. It was also reported that modafinil exerts some effects on cocaine reward and reinforcement.[34] As a result, its potential utility to establish abstinence from cocaine addiction was tested in a phase I clinical trial, which reported a cocaine-blunting effect.[35] As an atypical DAT inhibitor, even with a DAT affinity (K_i = 8.16 μM[1]) lower than that of cocaine, modafinil could still prevent DAT from being blocked by cocaine.

Using the proteome-informed ML models, we also predicted the side effects of modafinil on other proteins in our cocaine addiction PPI network. Consistent with the fact that modafinil is an approved drug, its predicted BA values to these side-effect targets are all larger than −9 kcal/mol, and among them, 65.7% are larger than −8 kcal/mol. In particular, modafinil only very weakly binds to hERG (−6.96 kcal/mol).

Because modafinil is a potential treatment for cocaine addiction, many of its analogues have been developed and tested. The first modafinil analogue is JJC8-016.[36] Because it is a potent atypical DAT inhibitor (K_i = 0.116 μM), its pretreatment inhibits cocaine-enhanced locomotion, cocaine self-administration, and cocaine-induced reinstatement of drug-seeking behavior. However, an experimental study found that it may block the hERG channel.[37] Our prediction for JJC8-016 also showed a BA value of −10.34 kcal/mol for hERG, which suggests serious potential cardiotoxicity.

Another series of modafinil analogues consists of JJC8-088, JJC8-089, JJC8-091,[38] and RDS3-094.[39] Their K_i values to DAT are 0.0026, 0.0378, 0.23, and 0.0231 μM, respectively. Although JJC8-091 is not as potent as JJC8-088 and JJC8-089, its efficiency in blocking cocaine effects is the best among these three. The possible reason is that JJC8-091 prefers an inward-facing conformation of DAT and thus exhibits a stronger atypical inhibition.[38] As a result, JJC8-091 is under further investigation. RDS3-094 was identified as a newly developed modafinil analogue in 2020. With similar chemical structures, JJC8-088, JJC8-089, JJC8-091, and RDS3-094 were predicted to have similar side effects. In our prediction, they all strongly bind to targets Sigma1 (BA values of −10.06, −10.15, −10.16, and −9.99 kcal/mol, respectively) and YES1 (BA values of −9.15, −9.31, −9.25, and −9.14 kcal/mol, respectively), and YES1 is a risky target associated with sarcoma. Additionally, JJC8-088 strongly interacts with SSTR5, which is related to pituitary adenoma. In contrast, their hERG side effects are not very obvious with BA values of −8.59, −8.48, −8.03, and −8.08 kcal/mol, respectively. Interestingly, these four modafinil analogues were also predicted to inhibit D_3R (−9.51, −9.07, −9.10, and −9.04 kcal/mol, respectively), which suggests they may assuage cocaine addiction via multiple mechanisms.

Benzatropine is a medication for movement disorders, including dystonia and parkinsonism. Benzatropine and its analogues are being studied for repurposing against cocaine addiction. Although benztropine pretreatment failed to significantly affect responses to acute cocaine administration,[40] its analogues are still worth investigating. For instance, JHW007, with a K_i of 0.0253 μM, shows antagonism of behaviors produced by cocaine or methamphetamine across numerous animal models.[41] However, serious side effects of JHW007 were predicted on targets Sigma1, GRM2, and YES1 with affinities of −9.46, −9.30, and −9.28 kcal/mol, respectively. The abnormality of YES1 could lead to sarcoma.

Rimcazole was originally designed as a potential antipsychotic. However, trials indicate rimcazole is not effective in this application. Instead, rimcazole and its analogues can reduce the effects of cocaine,[42] specifically through binding to DAT. Rimcazole was reported to have a K_i of 0.0977 μM to DAT.[1] Because it is still a drug candidate under investigation, the side effects remain a major concern. Our results indicate potentially serious side effects on targets D_4R and YES1 with BA values of −9.60 and −9.24 kcal/mol, respectively.

Rimcazole analogue GBR12909 (vanoxerine) is a potent atypical DAT inhibitor with a K_i of 0.00177 μM. Therefore, it was advanced to phase I clinical trials. However, a failure was reported due to rate-dependent corrected QT (QTc) elongation in healthy subjects.[1] This heart-related side effect was supported by the BA value for hERG, which was predicted to be as low as −9.47 kcal/mol. Another potential side effect is from Sigma1, with a predicted BA value of −9.43 kcal/mol.

Dextroamphetamine is an approved drug prescribed for treating ADHD and narcolepsy. Because it has potent atypical antagonism of DAT with a K_i of 0.109 μM,[43] it is being reexamined for the treatment of cocaine addiction. In clinical trials, dextroamphetamine has demonstrated significant promise for weakening the effects of cocaine.[44] Consistent with the fact dextroamphetamine is already approved, the predicted side effects on other targets are negligible in our models. For example, the predicted BA value versus hERG is −7.09 kcal/mol.

The D_3R antagonist SB277011A is studied for cocaine addiction. It can reduce cocaine-enhanced brain-stimulation reward, suppress cocaine-conditioned place preference, and attenuate cocaine-primed reinstatement behaviors in rats.[45] However, SB277011A is ineffective when cocaine is available for self-administration.[46] Considering side effects, SB277011A has approximately 100-fold selectivity for D_3R (K_i = 0.0112 μM) over D_2R (K_i = 555 μM). We also predicted its other side effects on more dopamine-related targets, and high predicted affinities occur on targets YES1, SSTR5, HGF, and LRRK2 (BA values of −9.71, −9.39, −9.23, and −9.17 kcal/mol, respectively). Among them, the abnormality of YES1 and SStR5 could trigger serious consequences because YES1 and SSTR5 are involved in sarcoma and prolactin-secreting pituitary adenoma.

NGB2904 is another widely investigated D_3R antagonist. In rats, NGB2904 reduces progressive ratio break points for cocaine, inhibits both cocaine- and cue-primed reinstatement of cocaine seeking, and prevents cocaine-enhanced brain-stimulation reward. However, its reward-attenuating effects can be overcome with larger doses of cocaine.[47] In our side-effect prediction, a potential side effect on hERG was detected with a predicted BA value of −9.20 kcal/mol, suggesting a risk of causing heart issues. Serious side effects were also predicted on multiple targets, including YES1, activin1, SSTR5, CNR1, and Sigma1 with BA values of −9.64, −9.60, −9.57, −9.53, and −9.47 kcal/mol, respectively, again raising serious concerns over sarcoma and prolactin-secreting pituitary adenoma. However, interestingly, NGB2904 was also predicted to potently inhibit DAT, suggesting NGB2904 may interact with multiple targets (D_3R and DAT) to affect cocaine addiction behaviors.

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

PG01037 is derived from NGB2904. Similar to NGB2904, PG01037 reduces progressive ratio break points for cocaine, inhibits both cocaine- and cue-primed reinstatement of cocaine seeking, and prevents cocaine-enhanced brain-stimulation reward but fails to alter cocaine self-administration.[48] PG01037 was predicted to have side effects that are weaker than those of NGB2904, especially as predicted BA values to hERG and YES1 are increased to −8.25 and −9.11 kcal/mol, respectively. Moreover, similar to NGB2904, PG01037 has a low predicted BA value versus DAT (−9.56 kcal/mol).

Many D_3R partial agonists are also currently under investigation. VK4-40 and VK4-116 are D3R partial agonists designed within the past five years. They were developed to improve blood−brain barrier (BBB) penetration of previous D_3R partial agonists. Compared with VK4-40, VK4-116 can reduce the cocaine-enhanced heart rate and blood pressure.[49] As for side effects, VK4-40 and VK4-116 exhibit high selectivity for D_3R (305- and 1735-fold more selective for D_3R over D_2R, respectively). However, our predictions show that they have potentially strong side effects on targets YES1 (BA values of −10.07 and −10.27 kcal/mol, respectively), SSTR5 (BA values of −9.44 and −9.66 kcal/mol, respectively), and LRRK2 (BA values of −9.28 and −9.30 kcal/mol, respectively).

The partial D_3R agonist CJB090 more effectively attenuated psychostimulant reward than PG01037 in rats.[50] We predicted its side effects on other targets, revealing the BA for YES1 (−9.27 kcal/mol) is weaker than those of VK4-40 and VK4-116. However, the potential side effect on SSTR5 is still somewhat high (BA value of −9.50 kcal/mol).

Cariprazine is an atypical antipsychotic already on the market. It acts primarily as a D_3R and D_2R partial agonist, with high selectivity for D_3R, and it is also being studied for the treatment of cocaine addiction. Cariprazine exhibits a reduction of cocaine intake under an FR1 schedule and reinstatement of cocaine seeking in rats.[51] It is only predicted to have a strong side effect (−9.82 kcal/mol) on Sigma1, and Sigma1 is not currently known to be related to serious diseases.

Other experimental D_3R partial agonists include BP-897, RGH-237, and GSK598809. BP-897 reduces conditioned locomotor activity to cocaine but fails when cocaine is self-administered.[53] As for side effects, the K_i assays indicate BP-897 is ∼70-fold more selective for D_3R over D_2R. Other side effects were predicted via our proteome-informed models. Potential serious side effects were found for targets SSTR5 and YES1 with predicted BA values of −9.73 and −9.38 kcal/mol, respectively. RGH-237 was also predicted to have somewhat high affinities for SSTR5 and YES1 (BA values of −9.79 and −9.65 kcal/mol, respectively). Notably, our predictions suggest GSK598809 strongly binds to targets YES1 and CNR1 with BA values of −10.15 and −9.94 kcal/mol, respectively, especially a −10.15 kcal/mol affinity for YES1 that may represent a danger for serious side effects, because YES1 is related to sarcoma.

The reuptake of dopamine from the synaptic cleft into the presynaptic axon terminal relies on a conformational cycle of DAT with five steps. The cycle begins with the ion/substrate-free (apo) state of DAT. This apo state is an outward-facing state in which DAT is open to the extracellular environment. In the first step, Na^+ binding stabilizes DAT in the fully outward-facing state with the extracellular gate entirely open, which is ready for the binding of dopamine to the primary S1 site. In the second step, dopamine binding induces closure of the extracellular gate, rendering the occluded conformation. In the third step, a second dopamine molecule binds to the S2 site in the extracellular vestibule and triggers the transition of DAT to the full inward-facing state open to the presynaptic axon terminal. As a result, dopamine and ions can be released into the presynaptic axon terminal. In the fourth step, after dopamine and ions are released, the apo inward-facing state of DAT is formed. In the final rate-limiting step, the DAT reverts to the outward-facing apo state, allowing the initiation of another translocation cycle.[15]

Cocaine and other typical DAT inhibitors prevent dopamine reuptake because the binding of typical inhibitors stabilizes the outward-facing state of DAT so that the intracellular gate is always closed and dopamine cannot be released into the cytosol, while atypical inhibition allows the transition to the inward-facing states of DAT, in which dopamine can gain access to the cytosol. The molecular mechanisms leading to typical or atypical inhibition can be unveiled via the comparison between the X-ray structure of cocaine bound to *Drosophila* DAT [Protein Data Bank (PDB) entry 4XP4 (see Figure 4a,b)] and that of 3,4-dichlorophenethylamine (DCP) with *Drosophila* DAT [PDB entry 4XPA (see Figure 4c)].[52] It suggests that the key to determining typical or atypical inhibition is the orientation of Phe319 in *Drosophila* DAT (corresponding to Phe320 in human DAT). Phe319 is the S1-gating residue, which is critical for the conformational transition from the outward-facing state to the inward-facing state of DAT. However, hindered by the bulky tropane ring and the methyl ester group present in cocaine, the side chain of Phe319 is stuck in the open state with the S1 pocket always open (see Phe319 in Figure 4b) so that the occluded and inward-facing states of DAT cannot be formed and the dopamine reuptake cycle cannot be completed. In contrast, the X-ray structure of DCP bound to *Drosophila* DAT shows an occluded state (Figure 4c). This is because, in this structure, DCP binding allows the side chain of Phe319 to rotate, occluding the S1 pocket (see Phe319 in Figure 4c), which is required by the following conformational movements in the dopamine reuptake cycle. Other research also suggests atypical inhibition accommodates the rotation of Phe320 in human DAT to the closed conformation.[38]

To confirm this atypical binding mechanism, we also performed induced-fit docking between the known atypical DAT inhibitor ibogaine and the occluded state of DAT (PDB entry 4XPA). Our induced-fit docking allows Phe319 to rotate freely. The resulting docking conformation suggests that ibogaine fits the occluded state of Phe319, which agrees with the atypical inhibition of ibogaine.

Here, using our proteome-informed ML models, we mimicked the processes of screening or repurposing lead compounds. We applied as many as 38 criteria in our systematic screening and repurposing. These criteria include BA for the designated target, six ADMET properties in Table 1, synthesizability, and the side effects on hERG and 30 other proteins related to cocaine addiction in Figure 2a.

Figure 4e exemplifies a nearly optimal lead from our systematic screening. In this example, we screened nearly optimal lead compounds from known DAT inhibitors in the ChEMBL database. We only accepted those with experimental DAT BA values of less than −9.54 kcal/mol (K_i < 0.1 μM), predicted hERG BA values of more than −8.18 kcal/mol (K_i > 1 μM), predicted BA values with other proteins of more than −9.54 kcal/mol, and excellent predicted ADMET properties

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

and synthesizability. As a result, screened compounds have high potencies to DAT, low side effects on hERG and other targets, and satisfy standards for druggable properties. Also, they are easy to synthesize. Compound CHEMBL558086 in Figure 4e was the only optimal lead compound left from our ML-base screening. It binds more strongly to DAT than to cocaine (experimental BA value of −11.32 kcal/mol) and has weak side effects on all of the other proteins, such as a predicted hERG BA value of −7.02 kcal/mol. The predicted ADMET properties and synthesizability are also in the excellent ranges (Table 1).

Panels f and g of Figure 4 show two example compounds from our systematic ML-based repurposing. In this experiment, we hope to repurpose inhibitors of PPI-informed proteins from ChEMBL to target DAT. For this purpose, we searched the compounds with weak affinity for their designated targets but potent binding to DAT by our ML predictions. At the same time, they must have weak side effects on other proteins such as hERG and good druggable properties. Our criteria are the same as those for the screening described above, but input compounds are from other data sets. As a result, the most potent DAT inhibitors from our systematic repurposing are ChEMBL3765422 and ChEMBL3764304 from the D_2R data set. They are very weak to D_2R, but they are predicted to be effective to DAT with BA values of −9.81 and −9.70 kcal/mol, respectively. Moreover, their side effects are weak. For example, its predicted hERG BA values are only −7.23 and −7.69 kcal/mol, respectively.

We also predict whether these three compounds are atypical or typical inhibitors via induced-fit docking. Our docking studies indicate that compound CHEMBL558086 from our systematic screening could bind to the occluded state of DAT with Phe319 closed, which suggests CHEMBL558086 is an atypical nearly optimal lead. Nevertheless, ChEMBL3765422 and ChEMBL3764304 from systematic repurposing fail to bind to the occluded state of DAT, and only the outward-facing state could accommodate them, suggesting typical inhibition by ChEMBL3765422 and ChEMBL3764304.

At present, there are no FDA-approved drugs for cocaine dependence. This work addresses the urgent need for effective drugs to treat this disease. We propose a proteome-informed machine learning platform, which results in 141 drug targets for cocaine dependence. Using our autoencoder trained from more than 104 million molecules, we build 32 machine learning models for the targets with enough existing training data. Using these models, we perform cross-target analysis of more than 60,000 drug candidates or experimental drugs to predict their side effects and repurposing potentials for treating cocaine addiction. We further screen the ADMET properties of these candidates and experimental drugs. Our platform reveals that essentially all existing drug candidates, including dozens of experimental drugs, fail to pass our cross-target and ADMET screenings, which explains why no FDA-approved anti-cocaine addiction drugs have yet to be uncovered despite decades of effort. Nonetheless, we have identified several nearly optimal leads for further optimization. Our induced-fit docking also suggests one atypical DAT inhibitor from them. Our work opens a novel, proteome-informed machine learning direction for drug discovery.

In this work, our PPI networks related to cocaine addiction were obtained from the String Web site (https://string-db.org/). The network analysis and visualization were implemented via Cytoscape 3.8.2.[54] In the network analysis, we considered three indices: degree, between centrality, and closeness centrality. The degree of a node is the number of edges. A node with a high degree represents a hub node having many neighbors. The between centrality of a node is defined as the proportion of the number of the shortest paths via it to the number of all of the shortest paths in the network, which quantifies the frequency at which a node forms the shortest paths between two other nodes. A node with a high between centrality is always a bottleneck of the network and dramatically influences the pathways among other nodes. Additionally, the closeness centrality of a node, which is defined as the average length of the shortest paths between the node and all other nodes, measures its centrality in the network. A node with a higher closeness centrality is closer to the center of the network.

Molecular fingerprints are the property profiles of a molecule, usually in the form of vectors with each vector element indicating the existence, degree, or frequency of one particular structure characteristic. They can be used as features for ML/DL models. In this work, the latent-vector fingerprint (LV-FP) and traditional 2D fingerprints (2D-FPs) were applied.

The seq2seq model is a DL autoencoder architecture that originated from natural language processing. It has already been demonstrated as a breakthrough success in English−French translation and conversational modeling. The basic scheme of the seq2seq model is to map an input sequence to a fixed-sized latent vector in the latent space using a gated recurrent unit (GRU)[55] or a long short-term memory (LSTM) network[56] and then to map the vector to a target sequence with another GRU or LSTM network.

In our study, input and output sequences are both SMILES strings, a one-dimensional "language" of chemical structures. Using nearly 104 million molecules, our autoencoder model is trained to have a high reconstruction ratio between input and output SMILES strings so that the latent vectors contain faithful information about the chemical structures. Thus, we adopted these latent vectors as LV-FP to represent compounds.

The seq2seq model and LV-FPs were realized by our in-house source code. We applied bidirectional LSTMs as the encoder. The generated LV-FP has a dimension of 512. The structure of our seq2seq model is illustrated in Figure S7.

In addition to LV-FP, 2D-FP-based predictors were also adopted in our work. The predictions from 2D-FPs were combined with those from LV-FP by consensus (averages of their prediction values) to further enhance the predictive power. According to our previous tests,[25] ECFP4,[57] Estate1,[58] and Estate2[58] fingerprints perform best on BA prediction tasks. Thus, these three 2D fingerprints were considered in this work. We employed the RDKit software (version 2018.09.3)[59] to generate 2D-FPs from SMILES strings.

To achieve fast and robust BA predictions, we performed gradient boosting decision tree (GBDT) regressors and classifiers using GradientBoostingRegressor and GradientBoostingClassifier modules in scikit-learn (version 0.20.1).[60] The hyperparameters were tuned according to 10-fold cross-validation tested on different data sets.

For regression tasks, the evaluation criteria are the square of the Pearson correlation coefficient (R^2) and the root-mean-square error (RMSE). For classification tasks, the evaluation criteria are accuracy, F score, sensitivity, and specificity.

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

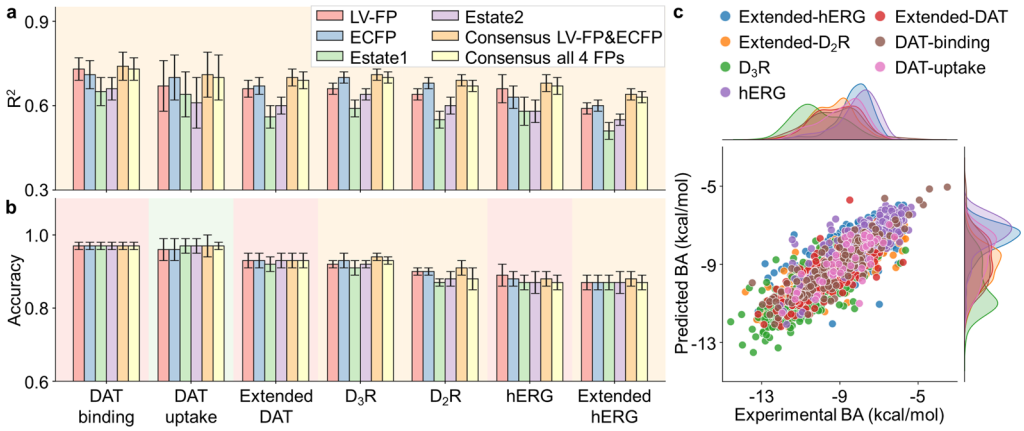


Figure 5. Performances of our predictive models using different fingerprints and their consensuses on some benchmark data sets relating to cocaine addiction. (a) Regression performances of our predictive models in terms of R^2. (b) Classification performance of our predictive models in terms of accuracy. (c) Comparison of our predicted binding affinities (BAs) from the LV-FP and ECFP consensus with the corresponding experimental BAs.

ADMET properties and synthesizability are also critical factors for drug design and lead optimization. In this work, we focused on seven different indices of ADMET and synthesizability: FDA maximum recommended daily dose (FDAMDD), log *P*, log *S*, half-life ($T_{1/2}$), Caco-2 permeability (Caco-2), human oral bioavailability 20% ($F_{20\%}$), and synthetic accessibility score (SAS). Among them, FDAMDD estimates the toxic dose threshold of a compound in humans. Log *P* is the logarithm of the *n*-octanol/water distribution coefficient. Log *S* is the logarithm of the aqueous solubility value. Log *P* and log *S* relate to the distribution of chemicals in human bodies.[61] The half-life of a drug indicates the length of time that the drug effect could persist in an individual. The value of $T_{1/2}$ here represents the probability of a half-life of <3 h. Caco-2 estimates *in vivo* drug permeability. $F_{20\%}$ measures the fraction of the initial dose of a drug that successfully reaches either the site of action or the bodily fluid domain from which the drug-intended targets have unimpeded access. Last but not least, SAS represents synthesizability, the ease of synthesizing a compound, based on a combination of fragment contributions and a complexity penalty. The score ranges from 1 (the easiest) to 10 (the hardest), which is calculated as a combination of two components: SAS = fragment score − complexity penalty.[62] We adopted ADMETlab 2.0 (https://admetmesh.scbdd.com/)[63] to predict these seven properties. Its document also provides optimal ranges for these ADMET properties, as shown in Table 1.

The induced-fit docking in our work was implemented via AutoDock Vina.[64]

This work studied 36 data sets for 32 different protein targets. The data were collected from refs 65 and 66 and the ChEMBL database.[29] These data sets are summarized in Table S1.

For each data set, we included all of the compounds with K_i or IC_{50} values but removed redundant ones. As suggested by Kalliokoski et al.,[67] the IC_{50} values were approximately converted to K_i by the equation $K_i = IC_{50}/2$. The label we used for training and testing is the binding affinity ($1.3633 \times \log_{10} K_i$).

In the benchmark classification tasks, the authors dropped the compounds between active and inactive to define a distinct boundary: for the DAT and hERG data sets, the compounds with pK_i or pIC_{50} values between 5 and 6 were excluded;[65] for the D_2R data set, the compounds with pK_i or pIC_{50} values between 6 and 7 were removed.[66] Therefore, to fairly compare our results to theirs, we also followed the same preprocessing strategies so that the data sets can be directly compared.

Some benchmark data sets related to cocaine addiction were already studied in terms of regression or classification tasks.[65,66] On these benchmark data sets, our predictors, especially consensus ones, exhibit high prediction power (see Figure 5). On regression tasks, we can achieve *R* values of >0.8 (R^2 > 0.64) on all of the data sets except extended hERG. The *R* on the extended hERG data set is close to 0.8 (R = 0.77; R^2 = 0.59). These performances are better than that in the existing report.[65] More detailed comparisons are shown in Tables S2–S13.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.1c03133.

Additional cross-target predictions, data sets and performance summary, and method and validation (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Guo-Wei Wei − *Department of Mathematics, Department of Biochemistry and Molecular Biology, and Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-5781-2937; Email: weig@msu.edu

Authors

Kaifu Gao − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0001-7574-4870

Dong Chen − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States*

Alfred J. Robison − *Department of Physiology, Michigan State University, East Lansing, Michigan 48824, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.1c03133

Notes

The authors declare no competing financial interest.

The 36 cocaine addiction-related data sets studied in this work are available at https://weilab.math.msu.edu/DataLibrary/2D. Our source code and trained autoencoder model for LV-FP generation can be found at https://github.com/WeilabMSU/antoencoder-v01.

■ ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health Grant GM126189, National Science Foundation Grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA Grant 80NSSC21M0023, the Michigan State University Foundation, Michigan Economic Development Corp., George Mason University Grant PD45722, Bristol-Myers Squibb 65109, and Pfizer. The authors thank Nancy Gilby for useful discussions.

■ REFERENCES

(1) Newman, A. H.; Ku, T.; Jordan, C. J.; Bonifazi, A.; Xi, Z.-X. New drugs, old targets: tweaking the dopamine system to treat psychostimulant use disorders. *Annu. Rev. Pharmacol. Toxicol.* **2021**, *61*, 609−628.

(2) Cheng, M. H.; Block, E.; Hu, F.; Cobanoglu, M. C.; Sorkin, A.; Bahar, I. Insights into the modulation of dopamine transporter function by amphetamine, orphenadrine, and cocaine binding. *Front. Neurol.* **2015**, *6*, 134.

(3) Matuskey, D.; Gallezot, J.-D.; Pittman, B.; Williams, W.; Wanyiri, J.; Gaiser, E.; Lee, D. E.; Hannestad, J.; Lim, K.; Zheng, M.-Q.; et al. Dopamine D3 receptor alterations in cocaine-dependent humans imaged with [11C](+) PHNO. *Drug Alcohol Depend.* **2014**, *139*, 100−105.

(4) Baik, J.-H. Dopamine signaling in reward-related behaviors. *Front. Neural Circuits* **2013**, *7*, 152.

(5) Paul, M.; Graybiel, A.; David, J.; Robertson, H. D1-like and D2-like dopamine receptors synergistically activate rotation and c-fos expression in the dopamine-depleted striatum in a rat model of Parkinson's disease. *J. Neurosci.* **1992**, *12*, 3729−3742.

(6) Ramanathan, S.; Irani, S. R. *Reference module in neuroscience and biobehavioral psychology*; Elsevier, 2018.

(7) Ptáček, R.; Kuželová, H.; Stefano, G. B. Dopamine D4 receptor gene DRD4 and its association with psychiatric disorders. *Med. Sci. Monit.* **2011**, *17*, RA215−220.

(8) Di Ciano, P.; Grandy, D. K.; Le Foll, B. Dopamine D4 receptors in psychostimulant addiction. *Adv. Pharmacol.* **2014**, *69*, 301−321.

(9) Unterwald, E. M. Regulation of opioid receptors by cocaine. *Ann. N. Y. Acad. Sci.* **2001**, *937*, 74−92.

(10) Rose, S. J.; Pack, T. F.; Peterson, S. M.; Payne, K.; Borrelli, E.; Caron, M. G. Engineered D2R variants reveal the balanced and biased contributions of G-protein and β-arrestin to dopamine-dependent functions. *Neuropsychopharmacology* **2018**, *43*, 1164−1173.

(11) Wu, B.; Hand, W.; Alexov, E. Opioid addiction and opioid receptor dimerization: structural modeling of the OPRD1 and OPRM1 heterodimer and its signaling pathways. *Int. J. Mol. Sci.* **2021**, *22*, 10290.

(12) Hope, B. T. Cocaine and the AP-1 transcription factor complex. *Ann. N. Y. Acad. Sci.* **1998**, *844*, 1−6.

(13) Robison, A. J.; Nestler, E. J. Transcriptional and epigenetic mechanisms of addiction. *Nat. Rev. Neurosci.* **2011**, *12*, 623−637.

(14) Schmitt, K. C.; Rothman, R. B.; Reith, M. E. Nonclassical pharmacology of the dopamine transporter: atypical inhibitors, allosteric modulators, and partial substrates. *J. Pharmacol. Exp. Ther.* **2013**, *346*, 2−10.

(15) Reith, M. E.; Blough, B. E.; Hong, W. C.; Jones, K. T.; Schmitt, K. C.; Baumann, M. H.; Partilla, J. S.; Rothman, R. B.; Katz, J. L. Behavioral, biological, and chemical perspectives on atypical agents targeting the dopamine transporter. *Drug Alcohol Depend.* **2015**, *147*, 1−19.

(16) Jordan, C. J.; Cao, J.; Newman, A. H.; Xi, Z.-X. Progress in agonist therapy for substance use disorders: Lessons learned from methadone and buprenorphine. *Neuropharmacology* **2019**, *158*, 107609.

(17) Hancox, J. C.; McPate, M. J.; El Harchi, A.; Zhang, Y. The hERG potassium channel and hERG screening for drug-induced torsades de pointes. *Pharmacol. Ther.* **2008**, *119*, 118−132.

(18) Cavalluzzi, M. M.; Imbrici, P.; Gualdani, R.; Stefanachi, A.; Mangiatordi, G. F.; Lentini, G.; Nicolotti, O. Human ether-à-go-go-related potassium channel: Exploring SAR to improve drug design. *Drug Discovery Today* **2020**, *25*, 344−366.

(19) Food and Drug Administration. International conference on harmonisation; guidance on S7B nonclinical evaluation of the potential for delayed ventricular repolarization (QT interval prolongation) by human pharmaceuticals; availability. Notice. *Fed. Regist.* **2005**, *70*, 61133−61134.

(20) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; et al. STRING v11: protein−protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607−D613.

(21) Ghislat, G.; Rahman, T.; Ballester, P. J. Recent progress on the prospective application of machine learning to structure-based virtual screening. *Curr. Opin. Chem. Biol.* **2021**, *65*, 28−34.

(22) Gao, K.; Nguyen, D. D.; Chen, J.; Wang, R.; Wei, G.-W. Repositioning of 8565 existing drugs for COVID-19. *J. Phys. Chem. Lett.* **2020**, *11*, 5373−5382.

(23) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative network complex for the automated generation of druglike molecules. *J. Chem. Inf. Model.* **2020**, *60*, 5682−5698.

(24) Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, 42717.

(25) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373−8390.

(26) Brown, J. M.; Hanson, G. R.; Fleckenstein, A. E. Regulation of the vesicular monoamine transporter-2: a novel mechanism for cocaine and other psychostimulants. *J. Pharmacol. Exp. Ther.* **2001**, *296*, 762−767.

(27) Mash, D. C.; Ouyang, Q.; Pablo, J.; Basile, M.; Izenwasser, S.; Lieberman, A.; Perrin, R. J. Cocaine abusers have an overexpression of α-synuclein in dopamine neurons. *J. Neurosci.* **2003**, *23*, 2564−2571.

(28) Vrana, S. L.; Vrana, K. E.; Koves, T. R.; Smith, J. E.; Dworkin, S. I. Chronic cocaine administration increases CNS tyrosine hydroxylase enzyme activity and mRNA levels and tryptophan hydroxylase enzyme activity levels. *J. Neurochem.* **1993**, *61*, 2262−2268.

(29) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(30) Flower, D. R. *Drug design: cutting edge approaches*; Royal Society of Chemistry, 2002.

(31) Smith, A. C.; Kupchik, Y. M.; Scofield, M. D.; Gipson, C. D.; Wiggins, A.; Thomas, C. A.; Kalivas, P. W. Synaptic plasticity mediating cocaine relapse requires matrix metalloproteinases. *Nat. Neurosci.* **2014**, *17*, 1655−1657.

(32) Efange, S. M.; Mash, D. C.; Khare, A. B.; Ouyang, Q. Modified ibogaine fragments: Synthesis and preliminary pharmacological characterization of 3-ethyl-5-phenyl-1, 2, 3, 4, 5, 6-hexahydroazepino [4, 5-b] benzothiophenes. *J. Med. Chem.* **1998**, *41*, 4486−4491.

(33) Koenig, X.; Hilber, K. The anti-addiction drug ibogaine and the heart: a delicate relation. *Molecules* **2015**, *20*, 2208−2228.

(34) Ballon, J. S.; Feifel, D. A systematic review of modafinil: potential clinical uses and mechanisms of action. *J. Clin. Psychiatry* **2006**, *67*, 554−566.

(35) Dackis, C. A.; Lynch, K. G.; Yu, E.; Samaha, F. F.; Kampman, K. M.; Cornish, J. W.; Rowan, A.; Poole, S.; White, L.; O'Brien, C. P.

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134

Modafinil and cocaine: a double-blind, placebo-controlled drug interaction study. *Drug Alcohol Depend.* **2003**, *70*, 29−37.
(36) Zhang, H.-Y.; Bi, G.-H.; Yang, H.-J.; He, Y.; Xue, G.; Cao, J.; Tanda, G.; Gardner, E. L.; Newman, A. H.; Xi, Z.-X. The novel modafinil analog, JJC8−016, as a potential cocaine abuse pharmacotherapeutic. *Neuropsychopharmacology* **2017**, *42*, 1871−1883.
(37) Tunstall, B. J.; Ho, C. P.; Cao, J.; Vendruscolo, J. C.; Schmeichel, B. E.; Slack, R. D.; Tanda, G.; Gadiano, A. J.; Rais, R.; Slusher, B. S.; et al. Atypical dopamine transporter inhibitors attenuate compulsive-like methamphetamine self-administration in rats. *Neuropharmacology* **2018**, *131*, 96−103.
(38) Newman, A. H.; Cao, J.; Keighron, J. D.; Jordan, C. J.; Bi, G.-H.; Liang, Y.; Abramyan, A. M.; Avelar, A. J.; Tschumi, C. W.; Beckstead, M. J.; et al. Translating the atypical dopamine uptake inhibitor hypothesis toward therapeutics for treatment of psychostimulant use disorders. *Neuropsychopharmacology* **2019**, *44*, 1435−1444.
(39) Slack, R. D.; Ku, T. C.; Cao, J.; Giancola, J. B.; Bonifazi, A.; Loland, C. J.; Gadiano, A.; Lam, J.; Rais, R.; Slusher, B. S.; et al. Structure−activity relationships for a series of (bis (4-fluorophenyl) methyl) sulfinyl alkyl alicyclic amines at the dopamine transporter: functionalizing the terminal nitrogen affects affinity, selectivity, and metabolic stability. *J. Med. Chem.* **2020**, *63*, 2343−2357.
(40) Penetar, D. M.; Looby, A. R.; Su, Z.; Lundahl, L. H.; Erös-Sarnyai, M.; McNeil, J. F.; Lukas, S. E. Benztropine pretreatment does not affect responses to acute cocaine administration in human volunteers. *Hum. Psychopharmacol.* **2006**, *21*, 549−559.
(41) Velázquez-Sánchez, C.; García-Verdugo, J. M.; Murga, J.; Canales, J. J. The atypical dopamine transport inhibitor, JHW 007, prevents amphetamine-induced sensitization and synaptic reorganization within the nucleus accumbens. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **2013**, *44*, 73−80.
(42) Katz, J. L.; Libby, T. A.; Kopajtic, T.; Husbands, S. M.; Newman, A. H. Behavioral effects of rimcazole analogues alone and in combination with cocaine. *Eur. J. Pharmacol.* **2003**, *468*, 109−119.
(43) Arunotayanun, W.; Dalley, J. W.; Huang, X.-P.; Setola, V.; Treble, R.; Iversen, L.; Roth, B. L.; Gibbons, S. An analysis of the synthetic tryptamines AMT and 5-MeO-DALT: emerging "novel psychoactive drugs. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 3411−3415.
(44) Grabowski, J.; Rhoades, H.; Schmitz, J.; Stotts, A.; Daruzska, L. A.; Creson, D.; Moeller, F. G. Dextroamphetamine for cocaine-dependence treatment: a double-blind randomized clinical trial. *J. Clin. Psychopharmacol.* **2001**, *21*, 522−526.
(45) Guerrero-Bautista, R.; Do Couto, B. R.; Hidalgo, J. M.; Cárceles-Moreno, F. J.; Molina, G.; Laorden, M. L.; Núñez, C.; Milanés, M. V. Modulation of stress-and cocaine prime-induced reinstatement of conditioned place preference after memory extinction through dopamine D3 receptor. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **2019**, *92*, 308−320.
(46) Di Ciano, P.; Underwood, R. J.; Hagan, J. J.; Everitt, B. J. Attenuation of cue-controlled cocaine-seeking by a selective D3 dopamine receptor antagonist SB-277011-A. *Neuropsychopharmacology* **2003**, *28*, 329−338.
(47) Xi, Z.-X.; Gardner, E. L. Pharmacological actions of NGB 2904, a selective dopamine D3 receptor antagonist, in animal models of drug addiction. *CNS Drug Rev.* **2007**, *13*, 240−259.
(48) Higley, A. E.; Spiller, K.; Grundt, P.; Newman, A. H.; Kiefer, S. W.; Xi, Z.-X.; Gardner, E. L. PG01037, a novel dopamine D3 receptor antagonist, inhibits the effects of methamphetamine in rats. *J. Psychopharmacol.* **2011**, *25*, 263−273.
(49) Shaik, A. B.; Kumar, V.; Bonifazi, A.; Guerrero, A. M.; Cemaj, S. L.; Gadiano, A.; Lam, J.; Xi, Z.-X.; Rais, R.; Slusher, B. S.; et al. Investigation of novel primary and secondary pharmacophores and 3-substitution in the linking chain of a series of highly selective and bitopic dopamine D3 receptor antagonists and partial agonists. *J. Med. Chem.* **2019**, *62*, 9061−9077.
(50) Orio, L.; Wee, S.; Newman, A. H.; Pulvirenti, L.; Koob, G. F. The dopamine D3 receptor partial agonist CJB090 and antagonist PG01037 decrease progressive ratio responding for methamphetamine in rats with extended-access. *Addict. Biol.* **2010**, *15*, 312−323.
(51) Roman, V.; Gyertyan, I.; Saghy, K.; Kiss, B.; Szombathelyi, Z. Cariprazine (RGH-188), a D3-preferring dopamine D3/D2 receptor partial agonist antipsychotic candidate demonstrates anti-abuse potential in rats. *Psychopharmacology* **2013**, *226*, 285−293.
(52) Wang, K. H.; Penmatsa, A.; Gouaux, E. Neurotransmitter and psychostimulant recognition by the dopamine transporter. *Nature* **2015**, *521*, 322−327.
(53) Cervo, L.; Carnovali, F.; Stark, J.; Mennini, T. Cocaine-seeking behavior in response to drug-associated stimuli in rats: involvement of D3 and D2 dopamine receptors. *Neuropsychopharmacology* **2003**, *28*, 1150−1159.
(54) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498−2504.
(55) Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, 1406.1078.
(56) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput* **1997**, *9*, 1735−1780.
(57) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(58) Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(59) *RDKit: Open-source cheminformatics*; 2006.
(60) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.
(61) Di, L.; Kerns, E. H. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug Discovery Today* **2006**, *11*, 446−451.
(62) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.
(63) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.* **2021**, *49*, W5−W14.
(64) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(65) Lee, K. H.; Fant, A. D.; Guo, J.; Guan, A.; Jung, J.; Kudaibergenova, M.; Miranda, W. E.; Ku, T.; Cao, J.; Wacker, S.; et al. Toward Reducing hERG Affinities for DAT Inhibitors with a Combined Machine Learning and Molecular Modeling Approach. *J. Chem. Inf. Model.* **2021**, *61*, 4266−4279.
(66) Warszycki, D.; Struski, L.; Smieja, M.; Kafel, R.; Kurczab, R. Pharmacoprint: A combination of pharmacophore fingerprint and artificial intelligence as a tool for computeraided drug design. *J. Chem. Inf. Model.* **2021**, *61*, 5054.
(67) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of mixed IC 50 data−a statistical analysis. *PLoS One* **2013**, *8*, No. e61007.

https://doi.org/10.1021/acs.jpclett.1c03133
J. Phys. Chem. Lett. 2021, 12, 11122−11134