nature communications

Perspective

https://doi.org/10.1038/s41467-026-73410-8

Mapping the avoid-ome: a systematic open-science approach to predictive ADMET

Received: 9 January 2026

Accepted: 4 May 2026

Published online: 25 May 2026

Check for updates

James S. Fraser[1] ✉, Steven Edgar[2], L. Naomi Handly[2], Sriram Kosuri[2], John D. Chodera[3], Mark Murcko[4] & W. Patrick Walters[5] ✉

Drug discovery often fails due to unpredictable ADMET issues, which account for 30% of clinical setbacks. Conventional methods lack the atomistic detail needed to navigate the "Avoid-ome"—a finite set of proteins acting as "anti-targets". OpenADMET is an open-science initiative addressing this by creating pre-competitive, mechanistic datasets. Using high-throughput structural biology, active learning, and community challenges, it builds generalizable models grounded in structural "ground truth". By directly studying the Avoid-ome, OpenADMET facilitates an era of rational, multi-parameter drug design.

Over the last 20 years, the number of new drug modalities have multiplied quickly[1]. Despite this, the oldest modality, small molecules, still accounts for ~3/4 of drugs approved by the FDA over the last decade[2,3]. Small molecules continue to dominate and even flourish for three main reasons: (1) small molecules have inherent distribution advantages with potential to reach every organ, cell, and organelle in the body; (2) modern engineering principles enable us to scale production and deliver them economically worldwide in predictable ways; and (3) our increased understanding of how small molecules engage proteins has led to a renaissance of new small molecule targeting modalities, such as covalent modifiers, correctors, allosteric modulators, induced proximity, degraders, and RNA/splicing/PPI/condensate modulators.

However, the ability to modulate all functions in the body also belies the challenges of dialing in unpredictable pharmacokinetic and safety properties. Drug discoverers must optimize for high potency at the target while avoiding interactions with related or idiosyncratic off-target proteins. The discovery process must also navigate problematic ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) issues that would cause a candidate to fail in preclinical development or, worse yet, clinical trials. Thus, drug design requires a multi-parameter optimization process that balances many different factors[1,3]. While every new target and related off-targets are unique to every program, the ADMET properties are shared, and thus, the focused development of predictive tools would broadly enable small-molecule drug discovery. Over the last 40 years, we've come to understand ADMET issues are largely driven by a finite set of proteins and physicochemical properties that can be measured in individual assays. While pharmacophore models and heuristics have been created to tackle specific liabilities, the field has not taken a systematic approach to understanding and correcting ADMET liabilities. Additionally, structure-based design, which has become a key part of modern drug discovery, is seldom used beyond primary target binding.

Despite significant progress in structure-based design and affinity prediction, ADMET properties remain the main reason for failure in drug discovery. More than 90% of molecules created during discovery fail to meet basic ADME standards[3–5]. Additionally, it is estimated that about 20% of drug candidates fail in preclinical toxicity tests, and around 30% of clinical failures are due to unexpected ADMET problems[3–5]. Traditional methods that focus mainly on bulk molecular properties[6], such as logP, solubility, or hydrogen bond donor counts, offer only vague guidance because they lack insight into the atomic-level interactions between drugs and the body's complex systems. Several machine learning (ML) representations[7], including chemical fingerprints, molecular graphs, 3D geometry-based models, and protein language models, are already aiding ADMET predictions. However, two key challenges hinder breakthroughs in predictive ADMET: the data remain extremely limited, and most models lack the atomistic detail needed for mechanistic understanding. What is missing from this perspective is a systematic, detailed understanding of the "Avoid-ome": the broad set of proteins that influence the ADME and toxicity properties of all drug molecules[8]. By focusing the tools of modern structure-based drug design on the Avoid-ome, drug discovery practice could be transformed.

[1]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA and Quantitative Biosciences Institute, University of California San Francisco, San Francisco, California, USA. [2]Octant, Inc, Emeryville, CA, USA. [3]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [4]Disruptive Biomedical LLC, Holliston, MA, USA. [5]OpenADMET, Boston, MA, USA. ✉e-mail: jfraser@fraserlab.com; pat.walters@omsf.io

Defining the avoid-ome

The Avoid-ome consists of enzymes, transporters, receptors, and channels that determine whether a compound reaches its intended target or fails due to off-target binding[9,10]. These include metabolic enzymes like cytochrome P450s (CYPs)[11], aldehyde oxidase[12], UDP-glucuronosyltransferases (UGTs)[13], and glutathione S-transferases (GSTs)[14]; transporters from the ABC[15] and SLC[16] families; plasma proteins such as serum albumin[17]; xenobiotic sensors like the pregnane X receptor (PXR)[18] and constitutive androstane receptor (CAR)[19]; and common toxicity drivers such as the hERG potassium channel[20], the voltage-gated sodium channel NaV1.5[21], and L-type calcium channels[22]. The primary Avoid-ome targets can be divided into four groups: absorption and excretion (A/E), distribution (D), metabolism (M), and toxicity (T) (Fig. 1).

Importantly, the Avoid-ome is a finite set: although thousands of proteins exist in the human proteome, only on the order of 50-100 proteins occur with high frequency as mediating ADMET properties; considering less common situations, perhaps a few hundred proteins in total are responsible for the preponderance of the ADMET challenges faced by discovery teams. While this initial target list is not definitive and is intended to start a conversation among the community, this bounded scope makes the Avoid-ome problem tractable if we can systematically generate and share the correct data.

Some Avoid-ome proteins are exploited therapeutically, such as P-gp in cancer[23] or SGLT2 in diabetes[24]. However, in most cases, drugs must be engineered to avoid them. Avoid-ome proteins are therefore not generally "targets" (the intended binding partner of a drug) or "off-targets" (proteins related to the intended binding partner). Instead, they are "anti-targets" (proteins that must be considered as potential confounders in most drug development projects).

Consider an example of a kinase inhibitor to illustrate the distinctions between targets, off-targets, and anti-targets (Fig. 2). Cyclin-dependent kinase 2 (CDK2)[25] is a vital serine/threonine protein kinase that plays a crucial role in regulating the eukaryotic cell cycle, particularly during the transition from the G1 phase to DNA synthesis (S phase). Abnormal (uncontrolled) activity of the CDK2/Cyclin E complex is commonly observed in many human cancers. This leads to increased cell proliferation and genomic instability, making CDK2 an important target for anti-cancer therapies. The primary off-targets of CDK2 inhibitors are usually other members of the CDK family because of their high similarity in ATP-binding sites. Ensuring off-target selectivity during the design of CDK2 inhibitors is essential for the safety and effectiveness of the drug. Lack of selectivity was a major reason why first-generation CDK2 inhibitors failed in early clinical trials. Additionally, anti-target selectivity against proteins such as CYPs and PXR is vital to prevent drug interactions, and avoiding the hERG ion channel is crucial in designing CDK2 inhibitors to prevent serious cardiac side effects. As the focus shifts from targets to off-targets and anti-targets, the number of binding sites and interactions to consider increases dramatically. The task of optimization is further complicated by the promiscuity of anti-targets, which have evolved to recognize a wide range of xenobiotics. This highlights the central challenge: success requires simultaneous optimization against every anti-target, yet systematic data on these interactions are almost entirely lacking.

The case for openADMET

How can we enable breakthroughs in ADMET modeling? Deep learning approaches crave data[26], yet very little is known about the ADMET properties of drug-like compounds. Publicly accessible databases, such as ChEMBL[27] and the Therapeutics Data Commons[28], include ADMET datasets compiled from the literature. However, this data has often been extracted from dozens of papers, each using different experimental procedures. As highlighted in a recent paper by Landrum and Riniker[29], reported values in the literature are rarely consistent. The pharmaceutical industry could be a valuable source of ADMET data. In fact, important ADMET datasets are stored within pharmaceutical companies, and making this data publicly available would be beneficial. However, even full access to this data would still be insufficient. To gain a comprehensive understanding, we need to systematically examine the interactions between the most common Avoid-ome targets and diverse sets of ligands that broadly cover chemical space.

The convergence of multiple factors has created an opportunity to systematically improve ADMET optimization. Recent advances in structural biology have enabled higher-throughput data collection and will allow us to study the structures of Avoid-ome proteins on a much larger scale. Additionally, advances in mass spectrometry have made data collection more cost-effective. Finally, the availability of open-source machine learning models has provided the ability to link structural and assay data to develop new strategies for compound optimization.

OpenADMET (http://openadmet.org), a major international open-science initiative initially funded by ARPA-H and the Gates Foundation, aims to leverage these scientific advances and address the

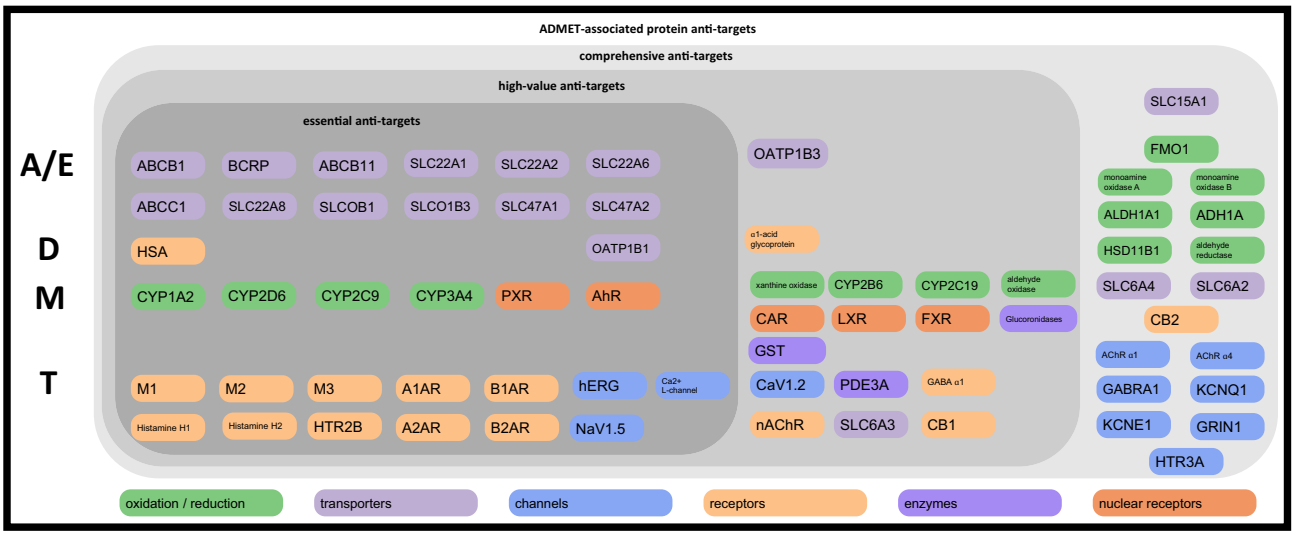

Fig. 1 | The set of protein anti-targets that comprise the Avoid-ome. These anti-targets can be categorized horizontally as essential, high-value, or comprehensive anti-targets. The colors in the plot further illustrate the mechanistic classes associated with each target.

https://doi.org/10.1038/s41467-026-73410-8

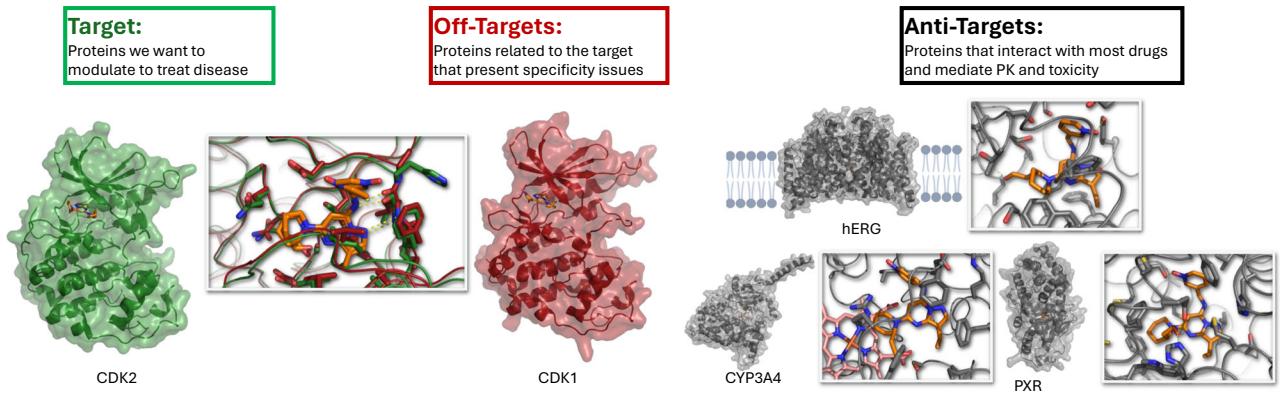

Fig. 2 | The distinction between targets, off-targets, and avoid-ome anti-targets. While on-target optimization usually involves interactions at a single binding site, considering off-targets introduces selectivity challenges as the major interactions driving affinity are likely conserved. Structures of an inhibitor bound in the active sites of CDK2 (PDB:4KD1) and CDK1 (PDB: 6GU6) are highly similar, with many interactions, such as the hinge binding motif highlighted by dashed lines, nearly identical. In contrast, structure predictions of the same compound bound to anti-targets such as hERG, CYP3A4, and PXR reveal a great diversity in hydrophobic and hydrogen bond interactions within the membrane protein hERG, the heme-containing binding pocket of CYP3A4, and the fully buried ligand binding site of PXR.

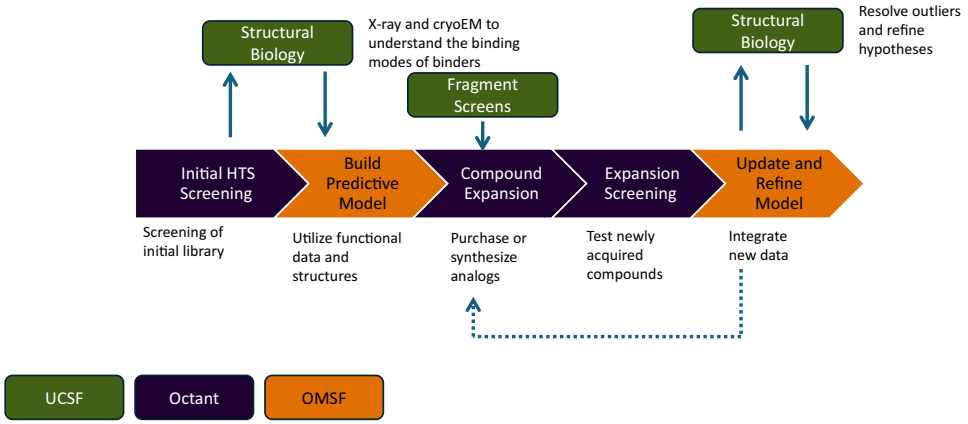

Fig. 3 | This schematic illustrates the collaborative pipeline employed by OpenADMET. While the process typically begins with initial HTS screening, it also allows for focused entry points via protein-ligand crystal structures to jump-start modeling or fragment screens to guide expansion. Integrated structural biology (X-ray/cryoEM) and functional data drive a continuous loop of model refinement and analog synthesis, shifting the focus from initial hits to exploring SAR (structure-activity relationship) around active compounds and resolving outliers.

ADMET data gap by developing pre-competitive, open datasets covering metabolism, transport, distribution, and toxicity. The OpenADMET consortium is a collaborative effort between the University of California, San Francisco (UCSF), Octant, and the Open Molecular Software Foundation (OMSF). We are creating platforms to make synthesizing compounds, conducting measurements, and learning from data cheaper and higher throughput. These large, openly accessible datasets will serve as a shared resource for the global research community. Our approach also leverages high-throughput structural biology to support model interpretability, identify outliers and cryptic binding modes, and ensure models are grounded in mechanistic, atomistic insights (Fig. 3).

In addition to the elements described above, genetic variation-aware predictions will be needed to realize the potential of pharmacogenomic analyses. We are integrating data such as VAMPseq[30] protein expression levels and atomistic modeling of how variants alter ADMET risk by changing interactions between the compound and the anti-target. Pre-clinical species translation is another essential component, requiring structural models that explain why compounds behave differently across rats, dogs, monkeys, and humans[31].

Another often-discussed alternative approach to feeding the data needs of machine learning approaches in ADMET is federated learning[32], where models are trained behind a firewall that can observe data from companies without sharing the underlying molecular identities. While appealing in principle, such solutions tend to reinforce limitations: they remain confined to local chemical space, they struggle to generalize, and they rarely deliver the mechanistic clarity required to address Avoid-ome proteins. By contrast, active learning[33] over diverse chemical space creates the conditions for truly generalizable models and deeper insights that can benefit the entire community. Freed from the constraints of having to generate molecules to "avoid" anti-targets, in OpenADMET we can synthesize and test compounds that are most informative for building predictive models. This close link between experiment and computation lets us ask, "which experiments will enable us to build the best model?"

Another key element of OpenADMET is the creation of blind community challenges to benchmark models using unreleased data, thereby promoting rigorous evaluation and continuous improvement. Collectively, these strategies will generate the data necessary for improved predictions and establish a framework for integrating functional data on specific anti-targets with bulk property assays, such

https://doi.org/10.1038/s41467-026-73410-8

	Nuclear Receptor Induction	CYP Inhibition	CYP Reactivity	GPCR Agonism & Antagonism	Microsomal Stability
Readout	Luminescence (Plate Reader)	Fluorescence (Plate Reader) Mass Spectrometry (Echo-MS+)	Mass Spectrometry (Echo-MS+)	Next Generation Sequencing	Mass Spectrometry (Echo-MS+)
Scale	1536-well plates 30K cmpds/run	1536-well plates 30K cmpds/run	1536 well plates 5K cmpds/run	384 well plates 10K cmpds/run	--
Measurements	EC50 - PXR , AHR KO - PXR, AHR	IC50 (TID & TDI) - CYP3A4, CYP1A2, CYP2D6, CYP2C9	Depletion & Cint - CYP3A4, CYP1A2, CYP2D6, CYP2C9	Agonism & Antagonsim for Aminergic GPCRs	--

Fig. 4 | The assays used to drive the current OpenADMET efforts. Additional assays will be added as new targets are introduced.

as stability in the presence of liver microsomes, to connect mechanistic understanding with applied pharmacology.

Assays

The OpenADMET initiative requires diverse assays across many anti-targets: biochemical assays for metabolism and transport, electrophysiology for ion channels, binding assays for plasma proteins, and transcriptional readouts for xenobiotic sensors (Fig. 4). A key goal is to modernize these assays, making them cheaper, more scalable, and more translatable. We begin by characterizing individual Avoid-ome targets to understand the mechanistic basis of ADMET liabilities before progressing to integrative assays like microsomal stability.

For metabolism and biochemical assays, we leverage scaled mass spectrometry[34] to enhance throughput and reduce costs. For cellular assays, we employ synthetic biology[35] to engineer high signal-to-noise genetic reporters and routinely include counter-assays with target knockouts to mitigate PAINS effects. Our goal is affordable, quantitative, scalable methods applicable to both purified compounds and those from next-generation direct-to-biology platforms. Currently, we evaluate CYP reactivity for thousands of compounds at <$0.40 per compound using an Echo-MS+ ZenoTOF 7600 system. Luminescence and fluorescence assays cost $0.05-$0.30 per well and can screen tens of thousands of compounds weekly.

Leveraging these high-throughput, cost-effective platforms enables us to rigorously interrogate the structure-activity relationship (SAR) landscape. Our initial screening establishes a baseline using a collection of marketed drugs alongside 10,000 diverse drug-like molecules. To explicitly identify and explore activity cliffs, which are highly valuable yet notoriously challenging to predict, we manage the scale of these experiments using an active learning workflow. This workflow incorporates a measure of model uncertainty[36,37] to mathematically balance exploration and exploitation, ensuring we only assay the most informative compounds to build generalizable models. For a concrete example of this workflow's scale, we typically begin by identifying approximately 10 commercially available analogs for each hit using an "analog by catalog" approach. When a specific property cliff requires deeper interrogation, we supplement these commercial sources with billion-scale synthesis-on-demand libraries or leverage our internal synthesis resources to design and rapidly generate focused sets of novel analogs to probe critical SAR drivers. Crucially, to maintain our high screening throughput and keep costs low during these focused iterations, our assays are robust enough to be conducted directly on crude reaction mixtures, completely bypassing time-consuming purification bottlenecks

While our framework heavily emphasizes the specific protein anti-targets of the Avoid-ome, we recognize that fundamental physicochemical and integrative properties—such as aqueous solubility, membrane permeability (logD), and metabolic stability are major drivers of ADMET outcomes, particularly for absorption and excretion. Although these factors are not mediated by a single anti-target, they are critical bulk molecular properties that often dictate whether a compound succeeds or fails. To ensure a comprehensive approach, OpenADMET is explicitly tackling these non-Avoid-ome properties alongside our protein-centric assays. This commitment was demonstrated by the very first OpenADMET/ExpansionRX community competition, which focused heavily on predicting these foundational physicochemical endpoints. Ultimately, our goal is to establish a unified framework that integrates bulk property assays, such as liver microsomal clearance, with precise functional data on specific Avoid-ome targets to bridge mechanistic understanding with applied pharmacology.

Chemistry

Our understanding of small-molecule Avoid-ome interactions is extremely sparse. There are, of course, existing examples of drugs known to interact with specific Avoid-ome targets. For example, terfenadine is a ~ 4 μM inhibitor of hERG[38] and ketoconazole inhibits CYP3A4 with an IC_{50} of ~40nM[39]. However, public data is usually not available for close analogs of these compounds. Furthermore, little is known about the Avoid-ome interactions that might exist within larger collections of drug-like molecules. To improve learning, we aim to quickly synthesize libraries of analogs to hits to identify activity cliffs[40], which are among the most valuable and challenging issues to predict over time. As with most drug discovery efforts, we have several methods to explore the structure-activity relationships (SAR) around compounds that bind to Avoid-ome targets. We can purchase compounds similar to known binders using an "analog by catalog" method, which has become a standard approach in pharmaceutical research. Starting with a library of 10,000 drug-like compounds, we typically have about 20 commercially available analogs for each one. These analogs can also be supplemented by synthesis-on-demand libraries[41], which currently number in the billions, or through custom synthesis at contract research organizations (CROs). An internal synthesis capability can provide an additional way to explore the SAR around Avoid-ome targets. By synthesizing and stockpiling key scaffolds and intermediates, hundreds of analogs can be rapidly generated and characterized. This process can potentially be further streamlined by avoiding time-consuming purification steps and conducting assays on crude reaction mixtures[42].

In addition to running assays on diverse chemical libraries, OpenADMET is also performing follow-up chemistry (Fig. 5) to further investigate SAR for Avoid-ome targets. The ADMET model-building process begins with screening commercial libraries or marketed drugs experimentally. After collecting an assay dataset, an active learning workflow can select additional compounds for synthesis and testing, which enhances the model's accuracy. This active learning approach employs a machine learning model that includes a measure of uncertainty, helping to balance exploration and exploitation during model

https://doi.org/10.1038/s41467-026-73410-8

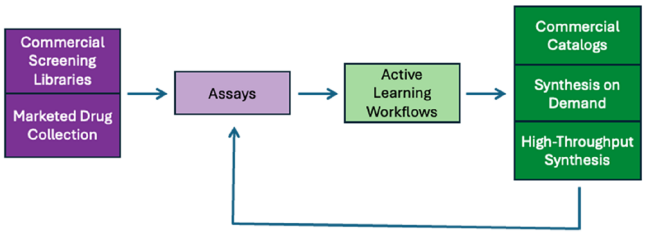

Fig. 5 | A diagram showing how chemistry is integrated into OpenADMET. After an initial assay is conducted, an active learning workflow is employed to select compounds for synthesis or purchase. This workflow combines exploration and exploitation to find additional compounds that will improve the model's performance.

to seed exploration of new regions of chemical space, whether by identifying new chemical matter and binding sites through X-ray fragment screening campaigns or by testing the outcomes of virtual screens. Determining the binding modes of compounds bound to Avoid-ome proteins is especially exciting from a structural and mechanistic perspective, precisely because they are promiscuous and often highly dynamic, able to accommodate diverse ligands in multiple poses. The promiscuity of Avoid-ome proteins makes them particularly difficult for co-folding methods (such as AlphaFold3) to discriminate binders from non-binders and to generalize to new chemical spaces with the correct binding poses. Our generated structures will therefore clarify the structural basis of outliers and activity cliffs identified by other models, providing the essential "ground truth" required to drive the field's predictive capabilities forward.

BOX 1

Some of the machine learning in drug discovery questions to be addressed by the OpenADMET effort

- What are the limits of an applicability domain, and can better metrics than simple similarity be defined?
- Do local models outperform global ones, and where can multitask models add value?
- What is the best way to fine-tune models, and do foundation models really help?
- How good is the data in the literature, and what are the best ways to split datasets?
- Is it realistic to generate enough high-quality experimental data at scale to train applicable models, given the large number of Avoid-ome targets, to prioritize the enormous chemical space involved?

development. The compounds chosen through active learning can then be synthesized using the various routes described above.

Structural biology's central role

Structural biology must also extend beyond traditional target enablement to illuminate the Avoid-ome[43]. Examples already include solving the structures of serum albumin complexes to explain drug partitioning, cryo-EM of transporters to decode efflux liabilities, and crystallography of PXR or CYP3A4 to guide metabolic risk mitigation. To support modeling efforts enabled by our larger chemical screening datasets, active compounds identified in our assays become prime candidates for structural elucidation. The number of structures generated will depend heavily on the target and experimental modality. For soluble targets, we can achieve a highly impactful scale; for example, we have already generated more than 100 X-ray structures of small molecules bound to PXR. For membrane proteins, the breakthrough capabilities of cryo-EM will provide a mechanistic basis for many transport and toxicity targets. While the throughput of our high-throughput chemical screening will naturally be much higher than that of X-ray crystallography, and even more dramatically than that of cryo-EM, we expect the throughput of structural biology to continue to increase.

Importantly, rather than training only our own predictive models, our goal is to empower the broader field to build multimodal models that integrate large-scale chemical screening data with this sparser, high-resolution protein-ligand interaction data. By determining the structures of active compounds and comparing them with inactive analogs, we will generate open data that enables the community to move beyond simple graph-based representations to scalable representations that accurately capture essential molecular interactions.

High-resolution structures provide critical insights, including snapshots of multiple protein and ligand conformations, a framework for modeling species differences and human genetic variants, and the grounding required for mechanistic understanding when machine learning predictions diverge. Structural biology is also uniquely suited

Computation

Over the past decade, there has been a notable increase in papers discussing the use of machine learning in chemistry and drug discovery. Unfortunately, the shortage of high-quality datasets has slowed progress in ADMET prediction. Most datasets used for training and benchmarking ADMET models are gathered from scientific literature. These datasets are often compiled from 20 to 50 different papers, each with distinct experimental conditions. Issues with reproducibility are demonstrated by a recent paper from Landrum and Riniker[29], where the authors compared IC_{50} values for the same compound tested against the same target in different papers. They found almost no correlation between the reported values. Given this variability, it is unrealistic to expect that datasets compiled from multiple literature sources will produce reliable machine learning models.

To build reliable models, we need large, consistently measured datasets of drug-like molecules that accurately reflect the data values observed in drug discovery. OpenADMET will create this data and share it with the community in a format suitable for training and evaluating machine learning models. This data will not only help train models but also promote the development of new molecular representations and algorithms. To advance the field, we must go beyond simple graph-based representations to scalable ones that accurately capture essential molecular interactions.

In addition to generating data, OpenADMET is developing a software framework, ANVIL, to record and codify best practices for machine learning model development. This open-source software allows users to easily test different molecular representations and machine learning algorithms. ANVIL also provides a range of tools for detailed evaluation and comparison of methods on high-quality datasets[44].

Community challenges and collaboration

Blind prediction challenges, inspired by CASP[45], CACHE[46], and SAMPL[47], will be a central component of OpenADMET. These challenges evaluate predictive models using unseen, high-quality data,

promote open-source methods and reproducibility, and maintain a shared scoreboard to track progress across academia and industry. Rather than viewing these challenges just as competitions, OpenADMET will use them as a catalyst to advance the current state of ADMET modeling. After each challenge, we will convene the community to discuss the most effective approaches and collaborate on computational and experimental strategies to drive further improvements.

The shortage of high-quality datasets has limited the field's ability to investigate key issues related to the application of ML models in ADMET prediction and drug discovery overall. Blind challenges, along with ML model development efforts within OpenADMET, provide a unique opportunity to explore fundamental questions about machine learning in ADMET; a sample of these questions is in Box 1. OpenADMET will generate the data needed to systematically study these questions.

Future perspective: from avoidance to design

The next decade of drug discovery will require shifting from ignoring Avoid-ome liabilities until late in drug development to embracing them early. Several broader questions will shape the field, especially as new modalities emerge. Are higher molecular weight molecules less prone to metabolism and transport because nature has not evolved to catch them? How specific are transporters – can molecules with structures quite different from the known substrates be transported? What is the Avoid-ome for antisense oligonucleotides? For proteins and peptides, at what molecular weight threshold does the body treat a drug as a peptide versus a protein? In each case, such questions can be addressed only by building and testing appropriate libraries.

There are also practical considerations for embedding Avoid-ome data into the discovery pipeline. For example, at what stage, hit-to-lead, lead optimization, or clinical candidate selection, should Avoid-ome predictive models or experiments be used either to triage ideas or to assist with multi-parameter optimization? Second, how can detailed Avoid-ome knowledge enable the exploitation of these anti-targets for the design of soft drugs or prodrugs? Third, can Avoid-ome efforts ever address unpredictable, rare toxicities like idiosyncratic DILI, which often only emerge late in Phase 3? Finally, how can interactions with "carrier" proteins such as serum albumin be optimized to extend drug half-life?

Conclusion

Understanding and navigating the Avoid-ome is the central universal challenge of modern drug discovery. By creating open, structural, and mechanistic datasets and benchmarking predictive models through blind challenges, OpenADMET provides a practical foundation for a new era of rational drug design. The best way to increase the effectiveness of drug discovery in the coming decade is to stop avoiding the Avoid-ome and instead study it directly.

References

1. Segall, M. D. Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* **18**, 1292–1310 (2012).
2. Murcko, M. A. What makes a great medicinal chemist? A personal perspective. *J. Med. Chem.* **61**, 7419–7424 (2018).
3. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–715 (2004).
4. Munson, M. et al. Lead optimization attrition analysis (LOAA): a novel and general methodology for medicinal chemistry. *Drug Discov. Today* **20**, 978–987 (2015).
5. Roberts, R. A. et al. Reducing attrition in drug development: smart loading preclinical safety assessment. *Drug Discov. Today* **19**, 341–347 (2014).
6. Caminero Gomes Soares, A. et al. Absorption matters: a closer look at popular oral bioavailability rules for drug approvals. *Mol. Inform.* **42**, e202300115 (2023).
7. Chuang, K. V., Gunsalus, L. M. & Keiser, M. J. Learning molecular representations for medicinal chemistry. *J. Med. Chem.* **63**, 8705–8722 (2020).
8. Fraser, J. S. & Murcko, M. A. Structure is beauty, but not always truth. *Cell* **187**, 517–520 (2024).
9. Bowes, J. et al. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discov.* **11**, 909–922 (2012).
10. Whitebread, S. et al. Secondary pharmacology: screening and interpretation of off-target activities – focus on translation. *Drug Discov. Today* **1**, 11 (2016).
11. Denisov, I. G., Makris, T. M., Sligar, S. G. & Schlichting, I. Structure and chemistry of cytochrome P450. *Chem. Rev.* **105**, 2253–2277 (2005).
12. Manevski, N., King, L., Pitt, W. R., Lecomte, F. & Toselli, F. Metabolism by aldehyde oxidase: Drug design and complementary approaches to challenges in drug discovery. *J. Med. Chem.* **62**, 10955–10994 (2019).
13. Oda, S., Fukami, T., Yokoi, T. & Nakajima, M. A comprehensive review of UDP-glucuronosyltransferase and esterases for drug development. *Drug Metab. Pharmacokinet.* **30**, 30–51 (2015).
14. Aloke, C., Onisuru, O. O. & Achilonu, I. Glutathione S-transferase: A versatile and dynamic enzyme. *Biochem. Biophys. Res. Commun.* **734**, 150774 (2024).
15. Thomas, C. & Tampé, R. Structural and mechanistic principles of ABC transporters. *Annu. Rev. Biochem.* **89**, 605–636 (2020).
16. Lin, L., Yee, S. W., Kim, R. B. & Giacomini, K. M. SLC transporters as therapeutic targets: emerging opportunities. *Nat. Rev. Drug Discov.* **14**, 543–560 (2015).
17. Ashraf, S. et al. Unraveling the versatility of human serum albumin - A comprehensive review of its biological significance and therapeutic potential. *Curr. Res. Struct. Biol.* **6**, 100114 (2023).
18. Ramanjulu, J. M. et al. Overcoming the pregnane X receptor liability: Rational design to eliminate PXR-mediated CYP induction. *ACS Med. Chem. Lett.* **12**, 1396–1404 (2021).
19. Willson, T. M. & Kliewer, S. A. PXR, CAR and drug metabolism. *Nat. Rev. Drug Discov.* **1**, 259–266 (2002).
20. Garrido, A., Lepailleur, A., Mignani, S. M., Dallemagne, P. & Rochais, C. hERG toxicity assessment: Useful guidelines for drug design. *Eur. J. Med. Chem.* **195**, 112290 (2020).
21. Abriel, H. Cardiac sodium channel Na(v)1.5 and interacting proteins: Physiology and pathophysiology. *J. Mol. Cell. Cardiol.* **48**, 2–11 (2010).
22. Lipscombe, D., Helton, T. D. & Xu, W. L-type calcium channels: the low down. *J. Neurophysiol.* **92**, 2633–2641 (2004).
23. Robinson, K. & Tiriveedhi, V. Perplexing role of P-glycoprotein in tumor microenvironment. *Front. Oncol.* **10**, 265 (2020).
24. Hsia, D. S., Grove, O. & Cefalu, W. T. An update on sodium-glucose co-transporter-2 inhibitors for the treatment of diabetes mellitus. *Curr. Opin. Endocrinol. Diabetes Obes.* **24**, 73–79 (2017).
25. Zhang, M. et al. CDK inhibitors in cancer therapy, an overview of recent development. *Am. J. Cancer Res.* **11**, 1913–1935 (2021).
26. Chen, J. et al. Data scaling and generalization insights for medicinal chemistry deep learning models. *J. Chem. Inf. Model.* **65**, 5887–5898 (2025).
27. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2011).
28. Huang, K. et al. Artificial intelligence foundation for therapeutic science. *Nat. Chem. Biol.* **18**, 1033–1036 (2022).
29. Landrum, G. A. & Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).

30. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
31. Musther, H., Olivares-Morales, A., Hatley, O. J. D., Liu, B. & Rostami Hodjegan, A. Animal versus human oral drug bioavailability: do they correlate? *Eur. J. Pharm. Sci.* **57**, 280–291 (2014).
32. Heyndrickx, W. et al. MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.3c00799 (2023).
33. Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discov. Today Technol.* **32-33**, 73–79 (2019).
34. Fan, X., Jiao, B., Zhou, X., Zhang, W. & Ouyang, Z. Miniaturization of mass spectrometry systems: An overview of recent advancements and a perspective on future directions. *Anal. Chem.* **97**, 9111–9125 (2025).
35. Kain, S. R. & Ganguly, S. Overview of genetic reporter systems. *Curr. Protoc. Mol. Biol.* **Chapter 9**, Unit9.6 (2001).
36. Parrondo-Pizarro, R., Lanini, J. & Rodríguez-Pérez, R. Uncertainty quantification in molecular machine learning for property predictions under data shifts. *J. Chem. Inf. Model.* **66**, 923–935 (2026).
37. Khalil, B., Schweighofer, K., Dyubankova, N., van Westen, G. J. P. & van Vlijmen, H. Combining Bayesian and evidential uncertainty quantification for improved bioactivity modeling. *J. Chem. Inf. Model.* **65**, 13057–13069 (2025).
38. Thouta, S., Lo, G., Grajauskas, L. & Claydon, T. Investigating the state dependence of drug binding in hERG channels using a trapped-open channel phenotype. *Sci. Rep.* **8**, 4962 (2018).
39. Eagling, V. A., Tjia, J. F. & Back, D. J. Differential selectivity of cytochrome P450 inhibitors against probe substrates in human and rat liver microsomes. *Br. J. Clin. Pharmacol.* **45**, 107–114 (1998).
40. Guha, R. & Van Drie, J. H. Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **48**, 646–658 (2008).
41. Kuan, J., Radaeva, M., Avenido, A., Cherkasov, A. & Gentile, F. Keeping pace with the explosive growth of chemical libraries with structure-based virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **13**, https://doi.org/10.1002/wcms.1678 (2023).
42. Hendrick, C. E. et al. Direct-to-biology accelerates PROTAC synthesis and the evaluation of linker effects on permeability and degradation. *ACS Med. Chem. Lett.* **13**, 1182–1190 (2022).
43. Stoll, F., Göller, A. H. & Hillisch, A. Utility of protein structures in overcoming ADMET-related issues of drug-like compounds. *Drug Discov. Today* **16**, 530–538 (2011).
44. Ash, J. R. et al. Practically significant method comparison protocols for machine Learning in small molecule drug discovery. *J. Chem. Inf. Model.* **65**, 9398–9411 (2025).
45. Gilson, M. K. et al. Assessment of pharmaceutical protein-ligand pose and affinity predictions in CASP16. *Proteins* **94**, 249–266 (2026).
46. Ackloo, S. et al. CACHE (Critical Assessment of Computational Hit-finding Experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat Rev Chem* **6**, 287–295 (2022).
47. Amezcua, M., Setiadi, J., Ge, Y. & Mobley, D. L. An overview of the SAMPL8 host-guest binding challenge. *J. Comput. Aided Mol. Des.* **36**, 707–734 (2022).

Acknowledgements

The authors would like to acknowledge the contributions of past and present members of the OpenADMET team, as well as the community that participates in our blind challenges.

Author contributions

All authors (J.F., S.E., L.N.H., S.K., J.C., M.M., and W.P.W.) contributed to the writing, editing, and revising of the manuscript.

Funding

This publication was supported by the Advanced Research Projects Agency for Health (ARPA-H) under AVOID-OME, and Award Number 1AY1AX000035. The contents are those of the authors. They may not reflect the policies of the Department of Health and Human Services or the U.S. government. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Advanced Research Projects Agency for Health.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-73410-8.

Correspondence and requests for materials should be addressed to James S. Fraser or W. Patrick Walters.