Journal of Cheminformatics

RESEARCH

Open Access

Data mining of PubChem bioassay records reveals diverse OXPHOS inhibitory chemotypes as potential therapeutic agents against ovarian cancer

Sejal Sharma[1], Liping Feng[1,2], Nicha Boonpattrawong[1], Arvinder Kapur[1], Lisa Barroilhet[1], Manish S. Patankar[1*] and Spencer S. Ericksen[3*]

Abstract

Focused screening on target-prioritized compound sets can be an efficient alternative to high throughput screening (HTS). For most biomolecular targets, compound prioritization models depend on prior screening data or a target structure. For phenotypic or multi-protein pathway targets, it may not be clear which public assay records provide relevant data. The question also arises as to whether data collected from disparate assays might be usefully consolidated. Here, we report on the development and application of a data mining pipeline to examine these issues. To illustrate, we focus on identifying inhibitors of oxidative phosphorylation, a druggable metabolic process in epithelial ovarian tumors. The pipeline compiled 8415 available OXPHOS-related bioassays in the PubChem data repository involving 312,093 unique compound records. Application of PubChem assay activity annotations, PAINS (Pan Assay Interference Compounds), and Lipinski-like bioavailability filters yields 1852 putative OXPHOS-active compounds that fall into 464 clusters. These chemotypes are diverse but have relatively high hydrophobicity and molecular weight but lower complexity and drug-likeness. These chemotypes show a high abundance of bicyclic ring systems and oxygen containing functional groups including ketones, allylic oxides (alpha/beta unsaturated carbonyls), hydroxyl groups, and ethers. In contrast, amide and primary amine functional groups have a notably lower than random prevalence. UMAP representation of the chemical space shows strong divergence in the regions occupied by OXPHOS-inactive and -active compounds. Of the six compounds selected for biological testing, 4 showed statistically significant inhibition of electron transport in bioenergetics assays. Two of these four compounds, lacidipine and esbiothrin, increased in intracellular oxygen radicals (a major hallmark of most OXPHOS inhibitors) and decreased the viability of two ovarian cancer cell lines, ID8 and OVCAR5. Finally, data from the pipeline were used to train random forest and support vector classifiers that effectively prioritized OXPHOS inhibitory compounds within a held-out test set (ROCAUC 0.962 and 0.927, respectively) and on another set containing 44 documented OXPHOS inhibitors outside of the training set (ROCAUC 0.900 and 0.823). This prototype pipeline is extensible and could be adapted for focus screening on other phenotypic targets for which sufficient public data are available.

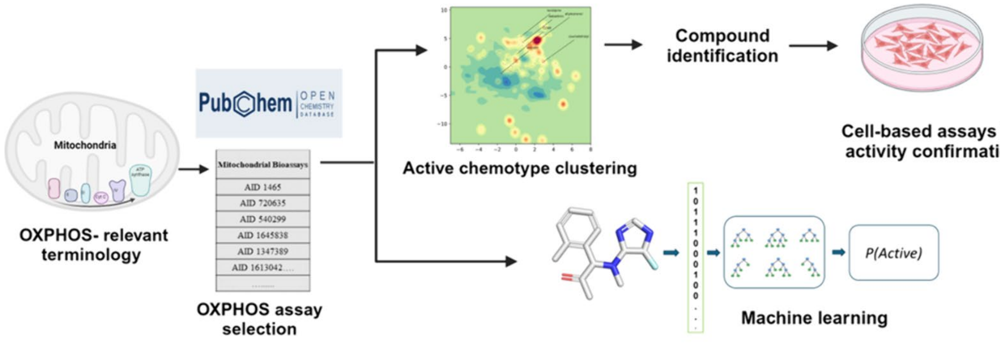
*Correspondence:
Manish S. Patankar
patankar@wisc.edu
Spencer S. Ericksen
sserickse n@wisc.edu
Full list of author information is available at the end of the article

Scientific contribution

Here, we describe and apply an assay data mining pipeline to compile, process, filter, and mine public bioassay data. We believe the procedure may be more broadly applied to guide compound selection in early-stage hit finding on novel multi-protein mechanistic or phenotypic targets. To demonstrate the utility of our approach, we apply a data mining strategy on a large set of public assay data to find drug-like molecules that inhibit oxidative phosphorylation (OXPHOS) as candidates for ovarian cancer therapies.

Keywords Data mining, PubChem, Oxidative phosphorylation, Electron transport chain, Ovarian cancer, Cheminformatics, Bioenergetics assays, Cancer therapeutics

Graphical Abstract



Created with BioRender.com

Introduction

The search for novel therapeutic compounds that target specific biological pathways remains a major undertaking in our quest for finding cures for many maladies. High throughput screening (HTS) remains a prominent approach for finding active compounds as starting points for probe or drug development. Given the resource commitment with HTS, investigators studying disease mechanisms have turned to smaller, focused screens using molecules prioritized by virtual screening (VS) models [1–3]. Both structure-based and ligand-based VS (LBVS) models continue to scale with rapidly expanding synthesize-on-demand virtual chemical libraries, now exceeding 40 billion molecules [4]. The LBVS models, however, have a distinct advantage in that they are broadly applicable to predicting any assay read-out including pathway or phenotypic end points where a specific protein target is unknown. The key limitation of such models, however, is that they may only be applied to targets for which ample molecule screening data are available for model training. For novel targets where such data are lacking, HTS assay data involving related protein targets might be used for training in a transfer learning-styled approach [5–7]. However, in such cases, it may not be clear which screening data sets should be included or whether data from a collection of varied assays can be effectively integrated for SAR-based data mining and model training. Here as a case study, we examine the utility in pooling disparate assay data sets associated with a multi-protein, anticancer target pathway to prioritize inhibitory chemical classes/scaffolds (chemotypes) and train LBVS models to guide focused screening.

There exists a plethora of pharmacological data on drug-like molecules in published literature from investigators around the world and through the last several decades (in some cases even the last century). With the establishment of PubChem [8], ChEMBL [9], BindingDB [10], and similar databases, vast chemical screening data have been collated and shared publicly. For example, PubChem, at the time of this writing, contains over 116 million compound records (CIDs) associated with >305 million bioassay results. We propose that careful selection and mining of such data can facilitate identification of drug-like molecules to test in focused screens. Here, we describe and apply an assay data mining pipeline to compile, process, filter, and mine public bioassay data. We believe the procedure may be more broadly applied to guide compound selection in early-stage hit finding on novel multi-protein mechanistic or phenotypic targets. To demonstrate the utility of our approach, we apply a data mining strategy on a large set of public assay data sets to find drug-like molecules that inhibit oxidative phosphorylation (OXPHOS). In the OXPHOS pathway, five multimeric protein complexes of the inner mitochondrial membrane cooperate to synthesize ATP. Inhibition of OXPHOS often results in a rapid increase in intracellular oxygen radicals, which in cancer cells cause DNA damage and cell death. Major efforts are underway to develop novel OXPHOS inhibitors that are safe and effective for treatment of ovarian and other malignancies [11–13]. The multiple proteins required for the functioning of OXPHOS provide an excellent opportunity to test the utility of our data mining approach to identify new small molecule anti-cancer drugs and hence is the focus of the work presented here. While some OXPHOS inhibitors such as atovaquone have proven to be safe and are approved for use in humans, others such as phenformin and IACS10759 have shown high toxicity in clinical trials and hence are no longer under consideration as therapeutic agents [11, 14–16]. Our approach allows clustering of the potential OXPHOS inhibitors based on their chemotypes, providing an essential resource for identification of chemical features that are essential to consider when developing novel OXPHOS inhibitors that are both effective as well as safe to use in clinical settings.

Results

Data mining workflow for identification of novel OXPHOS inhibitors

In previous studies we identified several small molecule inhibitors of OXPHOS (Fig. 1A) that produce an oxidative stress condition that shrinks tumors without significant toxic effects in non-cancer cells [11, 12, 17–21]. These earlier efforts focused on a small set of α/β-unsaturated carbonyl-containing compounds that likely competitively inhibit ubiquinone interactions with ETC complexes or potentially engage in redox reactions to prevent effective transfer of electrons through the OXPHOS pathway in the mitochondria. To identify additional compounds and compound classes (chemotypes) that promote oxidative stress in cancer cells by inhibition of OXPHOS (Fig. 1A), we developed a novel data mining workflow (Fig. 1B).

First, a broad search was performed to identify potentially relevant AIDs from the PubChem database. The query returned 8,415 unique AIDs with matches to “electron transport chain,” “mitochondrial complex,” “mitochondrial respiratory chain,” or “mitochondrial membrane potential,” in their AID assay descriptions (Fig. 1B). The search was further limited to assays involving small molecule screens. Terms such as the tricarboxylic acid pathway were excluded to focus on assays directly associated with the electron transport chain (ETC) components of the OXPHOS pathway. The 8415 matching bioassay records contained a total of 390,407 individual molecule testing results involving 312,039 unique Compound record identifiers (CIDs). The PubChem SMILES for the 312,039 CIDs were downloaded and processed using RDKit to canonicalize, sanitize and strip counterions. These procedures reduced the set to 310,557 unique molecule structures (SMILES) with unique CIDs. Note that most compound testing records were concentrated in a small fraction of the total bioassay records. Only two AIDs contained more than 100,000 unique CIDs (AIDs 1465 and 540229). Just five AIDs had greater than 10,000 CIDs and 10 AIDs had greater than 100.

Given the large number of assay records returned in this initial search, a more stringent filter was applied to reduce the bioassay set to those most relevant to the OXPHOS pathway. Based on inspection of the assay titles, we developed a list of terms for a secondary filter consisting of 39 positive terms (involving electron transport in mitochondria) and 2 negative terms (related to the photosynthesis pathway) (Additional file 1). These terms were searched in the text of the assay title, description, and abstract of each AID. We anticipated that matches to these specific terms (or non-matches to the negative terms) would remove assays less directly associated with ETC mechanisms. These priority AIDs are referred to as “ETC-linked.” This secondary filter reduced the set to 8003 ETC-linked AIDs, involving 228,240 unique CIDs and 227,554 unique SMILES (Fig. 1B).

Given our goal of identifying drug-like OXPHOS inhibitory molecules as starting points for clinical development, the remaining compounds with non-drug-like characteristics were flagged. Both a Lipinski-like pharmacokinetics (PK) filter (MW > 200 g/mol, MolLogP < 5.8, total polar surface area < 150, number of heavy (non-hydrogen) atoms > 20, number of nitrogen or oxygen atoms > 0, and at least one ring) and PAINS (Pan Assay Interference Compounds) filter were applied.

The molecules were then labeled as OXPHOS-active or -inactive based on the PubChem activity outcome categories [22] within the 8003 ETC-linked bioassay records. Compounds were labeled as ‘active’ if they tested positive (PubChem Activity Outcome == ‘Active’) in at least one assay and the number of positive assay results was equal to or greater than the number of negative testing results. Inactive compounds were defined as those that tested negative in at least one assay (PubChem Activity Outcome == ‘Inactive’) and having no ‘active’, ‘inconclusive’, or ‘unspecified’ outcomes. All remaining compounds were designated “unlabeled” and considered inconclusive.

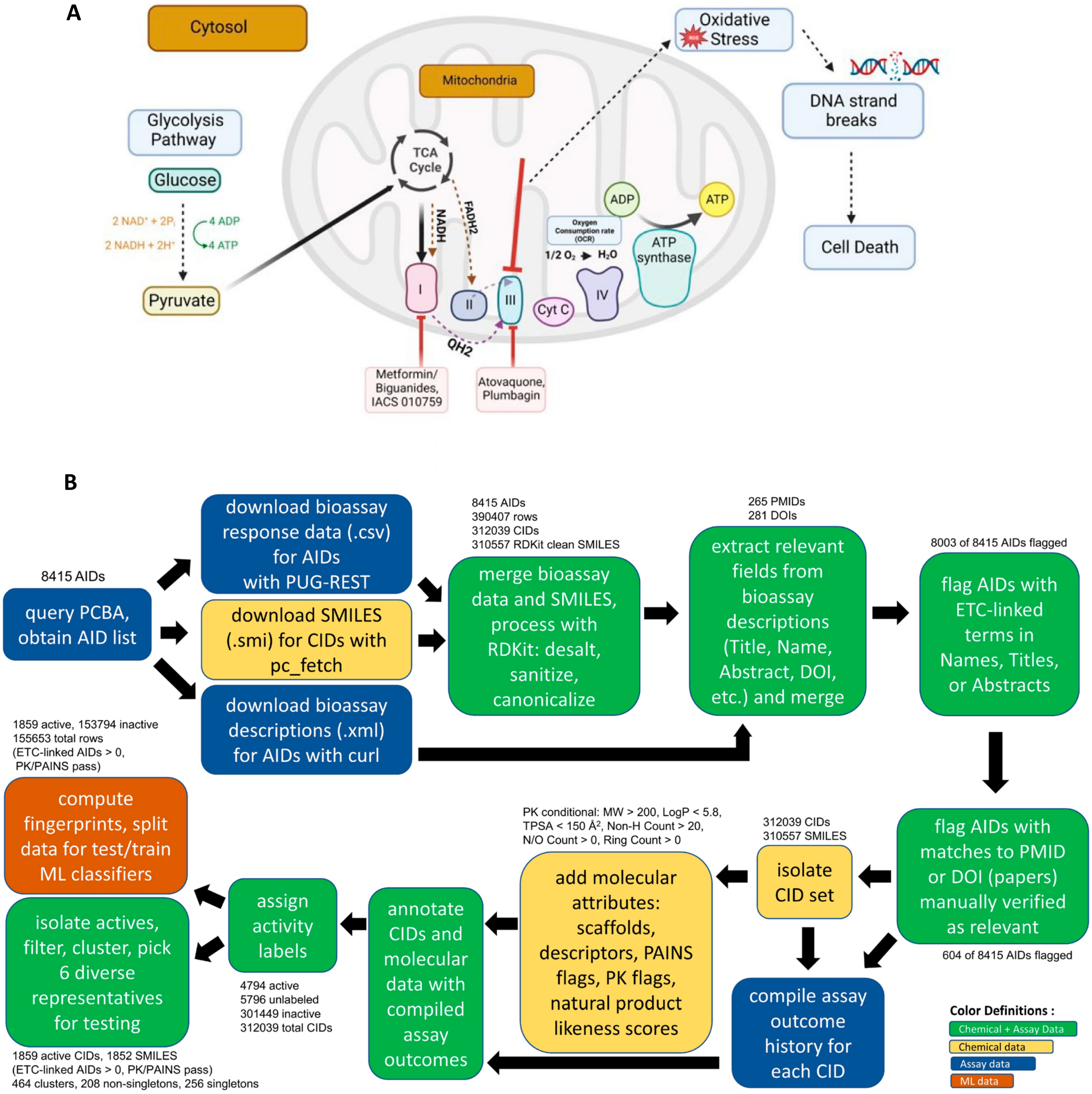

A
Cytosol
Mitochondria
Glycolysis Pathway
Glucose
2 NAD⁺ + 2Pᵢ
2 NADH + 2H⁺
4 ADP
4 ATP
Pyruvate
TCA Cycle
NADH
FADH2
I
II
III
QH2
Cyt C
IV
Oxygen Consumption rate (OCR)
1/2 O₂ → H₂O
ADP
ATP
ATP synthase
Metformin/ Biguanides, IACS 010759
Atovaquone, Plumbagin
Oxidative Stress
DNA strand breaks
Cell Death
B
8415 AIDs
query PCBA, obtain AID list
download bioassay response data (.csv) for AIDs with PUG-REST
download SMILES (.smi) for CIDs with pc_fetch
download bioassay descriptions (.xml) for AIDs with curl
8415 AIDs
390407 rows
312039 CIDs
310557 RDKit clean SMILES
merge bioassay data and SMILES, process with RDKit: desalt, sanitize, canonicalize
265 PMIDs
281 DOIs
extract relevant fields from bioassay descriptions (Title, Name, Abstract, DOI, etc.) and merge
8003 of 8415 AIDs flagged
flag AIDs with ETC-linked terms in Names, Titles, or Abstracts
flag AIDs with matches to PMID or DOI (papers) manually verified as relevant
604 of 8415 AIDs flagged
312039 CIDs
310557 SMILES
isolate CID set
compile assay outcome history for each CID
PK conditional: MW > 200, LogP < 5.8, TPSA < 150 Å², Non-H Count > 20, N/O Count > 0, Ring Count > 0
add molecular attributes: scaffolds, descriptors, PAINS flags, PK flags, natural product likeness scores
annotate CIDs and molecular data with compiled assay outcomes
assign activity labels
4794 active
5796 unlabeled
301449 inactive
312039 total CIDs
1859 active, 153794 inactive
155653 total rows
(ETC-linked AIDs > 0, PK/PAINS pass)
compute fingerprints, split data for test/train ML classifiers
isolate actives, filter, cluster, pick 6 diverse representatives for testing
1859 active CIDs, 1852 SMILES
(ETC-linked AIDs > 0, PK/PAINS pass)
464 clusters, 208 non-singletons, 256 singletons
Color Definitions :
🟩 Chemical + Assay Data
🟨 Chemical data
🟦 Assay data
🟧 ML data

Fig. 1 Data mining pipeline for identification of novel OXPHOS inhibitors. **A** The four multi-protein complexes (Complexes I–IV) of the OXPHOS pathway transfer of high energy electrons from NADH and $FADH_2$ to oxygen. The proton gradient in the intermembrane space during this electron transfer drives the synthesis of ATP by ATP synthase. Pharmacologic inhibitors of Complex I and III block electron transport resulting in oxidative damage of DNA and other biomolecules causing cell death. **B** Data mining pipeline for the identification of OXPHOS inhibitors is shown. A list of 8415 OXPHOS-related assay ID (AID) was identified through a search of the PubChem database. The pipeline was then used to filter out 1859 candidates as potential OXPHOS inhibitors from a total of 310,557 compounds

Based on our activity labeling, among the 227,554 unique molecules (228,240 CIDs) tested in the 8003 ETC-linked AIDs, there were 4140 active molecules (4199 CIDs). After applying PK and PAINS flags, the number of active molecules was reduced to 1852 (1859 CIDs). This "drug-like" active set was clustered based on structural similarity, yielding 464 unique chemical clusters (208 non-singletons). Structures of the medoids (most central compound in a chemical cluster) are shown in Additional file 2 to demonstrate their structural diversity. Additionally, the published records associated with these active compounds were identified. These records corresponded to articles with 74 unique PMIDs associated with active individual compounds from 297 AIDs. This result indicated that the algorithm used in our analyses was successful in reducing the set of relevant compounds (464 cluster representatives) by 670-fold from the original 390,407 compound testing records (312,039 CIDs and 310,557 unique compound structures) in the initial 8415 AIDs identified in the PubChem query.

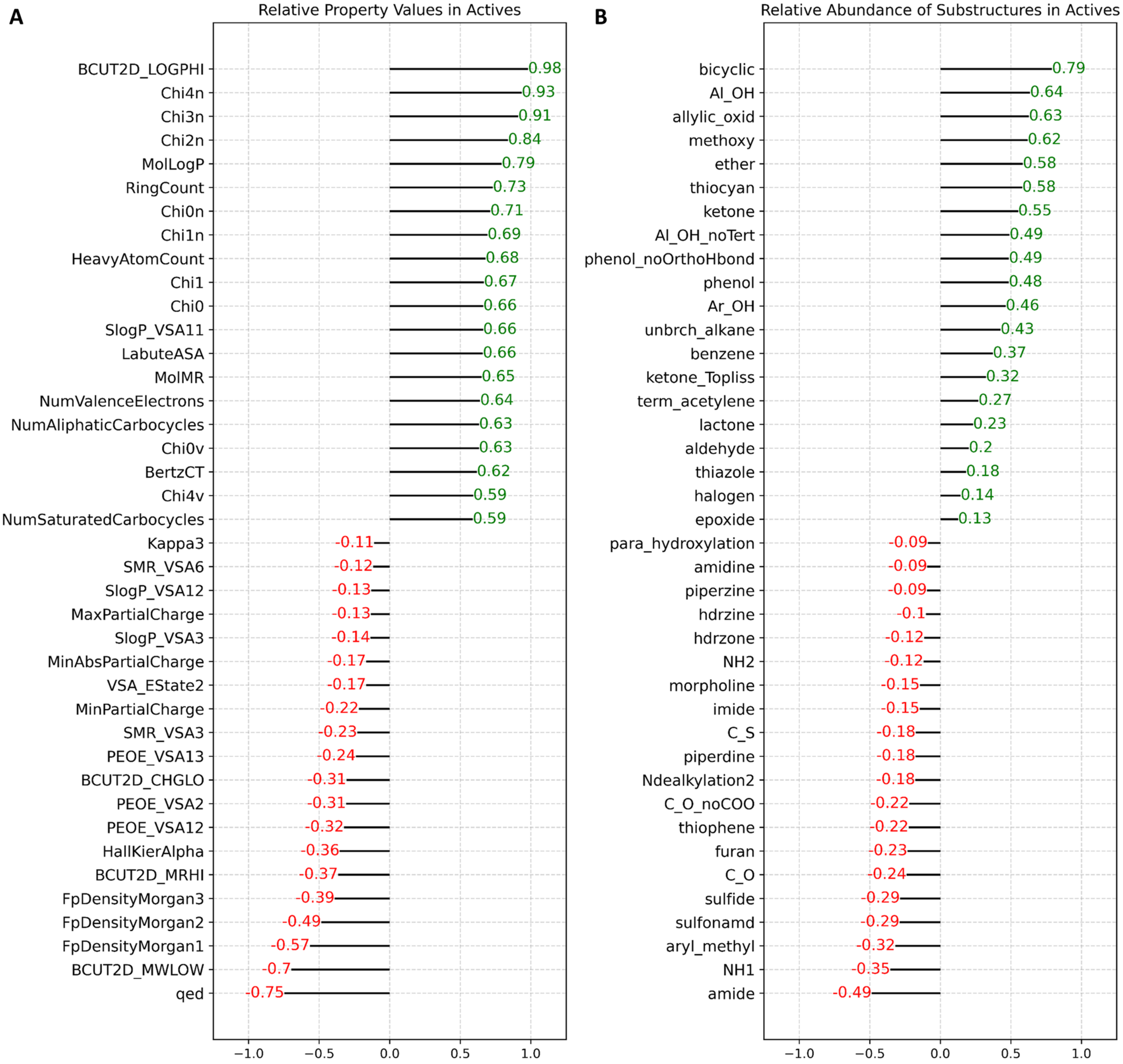


Fig. 2 Active compounds in the OXPHOS assay data set diverge from inactive compounds with respect to **A** chemical properties and **B** specific chemical substructures/functional groups. In **A** and **B**, the diverging bar plot shows the difference between active and inactive compound population means after transformation by standard scaling (mean = 0, standard deviation = 1). Difference bars are shown for only the top 20 and bottom 20 most skewed molecular attributes with respect to OXPHOS activity. Green values indicate shifts in favor of OXPHOS actives. Red values show shifts favoring OXPHOS inactives. Definitions of fragments and properties are provided in Additional file 8

Data mining

To identify distinguishing characteristics of OXPHOS inhibitor candidates, we compared both property and fragment/functional group distributions between the active and inactive compound populations (Fig. 2A and B). A subset of 155,653 unambiguously labeled molecules (153,794 inactive and 1859 active) derived from ETC-linked AIDs and that passed PK and PAINS filters were isolated for analyses. This set is referred to as the PC_OXPHOS training set. In terms of molecular properties, OXPHOS-active compounds exhibit elevated values for hydrophobicity-based features including BCUT2D_LOGPHI (highest eigenvalue for an atom adjacency matrix of elements as atom-wise logP contributions), MolLogP (predicted LogP), SlogP_VSA11, and MolMR (predicted molar refractivity). Active molecules also show relatively high values for size-related features like LabuteASA (van der Waals approximate surface area), number of atoms, molecular weight, number of rings, and aliphatic and aromatic carbocycles. OXPHOS actives also exhibit markedly lower Morgan fingerprint densities (molecular complexity) and quantitative estimate of drug-likeness (QED) values. Regarding structural characteristics, active compounds show higher abundance of bicyclics and oxygen-containing functional groups including ketones, allylic oxides (alpha/beta unsaturated carbonyls), hydroxyl groups (aliphatic, aromatic, and phenolic), and ethers. In contrast, amide and primary amine (NH1) functional groups show a notably lower than random prevalence.

UMAP analysis to identify chemotype clusters of OXPHOS inhibitor candidates

To assess the diversity of the OXPHOS active chemotypes and the extent to which they might be structurally distinguished from inactive compounds in a low dimensional chemical space, we performed a 2D Uniform Manifold Approximation and Projection (UMAP) on the PC_OXPHOS training set using Morgan fingerprint representations. Using the compounds' UMAP coordinates as data points, we plotted a contour map of the kernel density estimate (KDE) of the chemical space and then overlayed OXPHOS-active compounds as points on this space (Additional file 3A). Clusters larger than 2 compounds were colored with cluster medoids labeled by cluster ID numbers. All other active compounds (from doublets or singletons) were shown as unlabeled grey points. Contour maps of the KDE for OXPHOS-inactive (Additional file 3B) and OXPHOS-active (Additional file 3C) show substantial divergence in their locations within the chemical space. As might be expected with the marked class imbalance (1.2% active) in the PC_OXPHOS training set, inactive map looks nearly indistinguishable from the full compound distribution (Additional file 3A and B). However, the OXPHOS-active KDE shows many distinctive superclusters, groupings of related clusters, that depart from high-density regions of the general chemical space. The differences are more clearly illustrated in a difference map created by subtracting the inactive KDE from the active KDE (Fig. 3). The UMAP projection, a dimensionality reduction technique known to preserve local distance relationships, effectively co-localized cluster cohorts and mapped structurally related clusters into adjacent regions where fundamental chemotypes are shared within superclusters. For example, the highest density OXPHOS-active supercluster, consisting of clusters 1, 27, 28, 31, 37, 45, 46, 58, 65, and 66 (upper center-right of Fig. 3), mostly share a common steroid-like chemotype (a 4-ring poly-alicyclic scaffold) (Additional file 3D). Based on inspection of the UMAP projection, we categorized the OXPHOS-active densities into 3 broad chemical classes: (i) oxygen-rich, lightly unsaturated polycycles—which include the alpha/beta unsaturated groups, ketones, ethers, and alcohols (Additional file 3D), (ii) nitrogen-rich polycycles, including biguanide-like chemotypes with poly-nitrogenated rings, imine, and urea groups (Additional file 3E), and (iii) miscellaneous group that covers the remaining diverse structures (Additional file 3F). The oxygen-rich and nitrogen-rich groups share moieties with established OXPHOS-inhibitory agents like rotenone and metformin, respectively.

Testing candidate molecules selected from data mining for OXPHOS inhibitory activity

Next, we conducted targeted assays to verify the OXPHOS inhibitory activity of randomly selected candidate compounds identified through our data mining approach. Six compounds, oryzalin, allylestrenol, esbiothrin, lacidipine, coumatetralyl, and (Z)-2-cyano-*N*-cyclohexyl-3-(5-morpholin-4-ylfuran-2-yl) prop-2-enamide (CCMP), each representing a unique chemotype from a distinct location in active UMAP space (Fig. 3), were chosen to verify their OXPHOS inhibitory activities. All six compounds were purchased from established vendors as high purity (> 95%) formulations in DMSO.

OXPHOS inhibitory activity of the six test compounds was tested in bioenergetics assays conducted with the murine ovarian cancer cell line ID8 using the Seahorse XFe96 bioanalyzer. Cells were cultured in media with or without the drugs (0–40 μM) for 45 min prior to determining oxygen consumption rate (OCR), a measure of OXPHOS activity. Four of the six compounds tested, oryzalin, allylestrenol, esbiothrin, and lacidipine inhibited OCR in the ID8 cells (Fig. 4A–D). Coumatetralyl and CCMP did not inhibit OCR production in these assays indicating that they were not functioning as OXPHOS

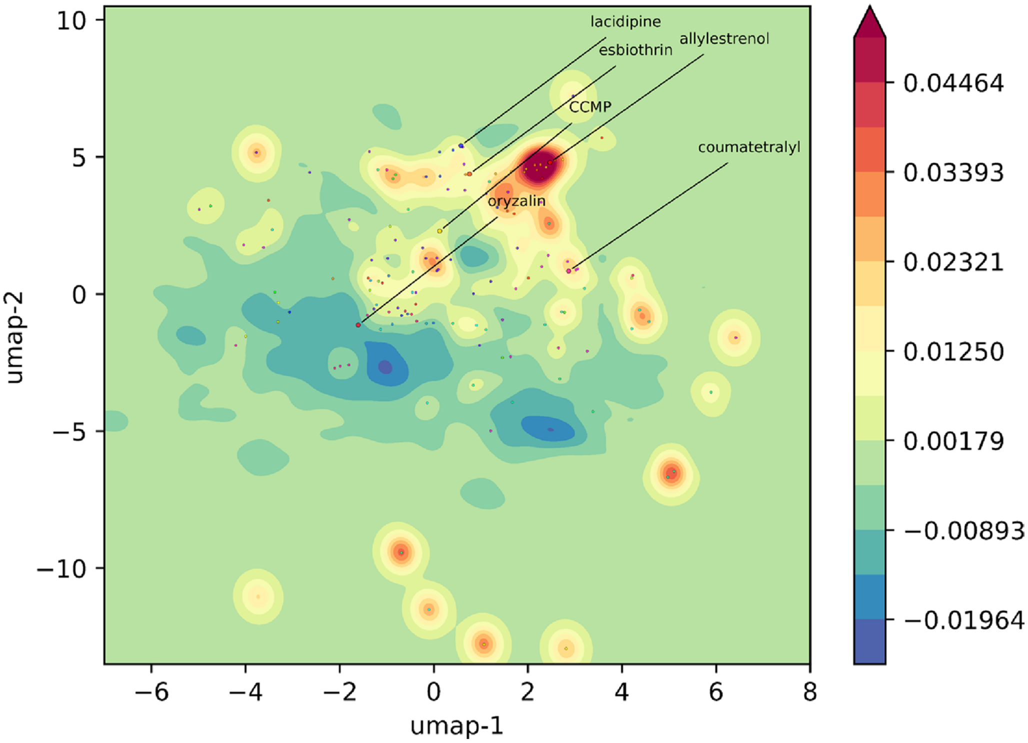


Fig. 3 UMAP showing density differences between active and inactive compound distributions in a 2-D chemical space of PC_OXPHOS—an unambiguously labeled subset of compounds tested in OXPHOS assays (PC_OXPHOS). The difference map was derived by subtracting the KDEs of the active and inactive points. Positive values (red) show regions of relatively high OXPHOS-active density. Negative values (blue) show relatively low density for OXPHOS actives. Medoids from active chemical clusters of size 3 or greater are overlayed as labeled points. The 6 compounds we acquired and tested on OXPHOS-related endpoints are also shown as points labeled by compound names

inhibitors (data not shown). In these experiments, lacidipine mediated the strongest dose-dependent inhibition of OCR in the ID8 cells (Fig. 4D). This decrease in OCR was complemented by inhibition of ATP production in the ID8 cells treated with oryzalin, allylestranol, esbiothrin, and lacidipine (Fig. 4A–D). Here again, among the six compounds tested, lacidipine was the strongest inhibitor of ATP production in ID8 cells (Fig. 4D).

In previous studies, we demonstrated that known OXPHOS inhibitors, atovaquone, plumbagin and others, by interfering with electron transport, induce a rapid intracellular oxygen radical flux [11, 12, 17, 18, 23]. This increase in the reactive oxygen radicals causes severe cellular damage leading to cell death. With four of the six test compounds demonstrating OXPHOS inhibitory activity, we next tested their ability to induce an intracellular oxygen radical surge using an established flow cytometry-based assay. The human and murine ovarian cancer cell lines, OVCAR-5 and ID8, respectively, were prelabeled with the oxygen radical sensing dye, H2DCFDA. Upon oxidation by reactive oxygen species, H2DCFDA is converted into green fluorescent DCF that is detected by flow cytometry. Using this assay, we noted that esbiothrin and lacidipine induced reactive oxygen flux in both OVCAR-5 and ID8 cells (Fig. 4E and F). While CCMP could not be tested in this assay because

(See figure on next page.)

Fig. 4 Oryzalin, allylestrenol, esbiothrin, and lacidipine inhibit OXPHOS and induce oxygen radical flux. **A**–**D**, ID8 cells preincubated with oryzalin, allylestrenol, esbiothrin and lacidipine (10–40 µM), respectively were examined in XFe96 Seahorse instrument for determination of oxygen consumption rate (OCR) and ATP production. DMSO was used as vehicle control in all experiments. Results shown above are the average of four technical replicates conducted with ID8 cell lines with $p < 0.05$ (**A**–**D**). **E**, **F**, ID8 and OVCAR-5 cells, respectively, were pre-labeled with the oxygen radical-sensing dye H2DCFDA were treated with oryzalin, allylestrenol, esbiothrin, lacidipine and coumatetralyl (10 µM) for 45 min. The increase in intracellular oxygen radicals was determined by flow cytometry (**E**, **F**). The bar graphs shown are representative of three biological replicates. The bar graphs show averages of the median fluorescence levels from the three technical replicates for each cell line. $p < 0.05$

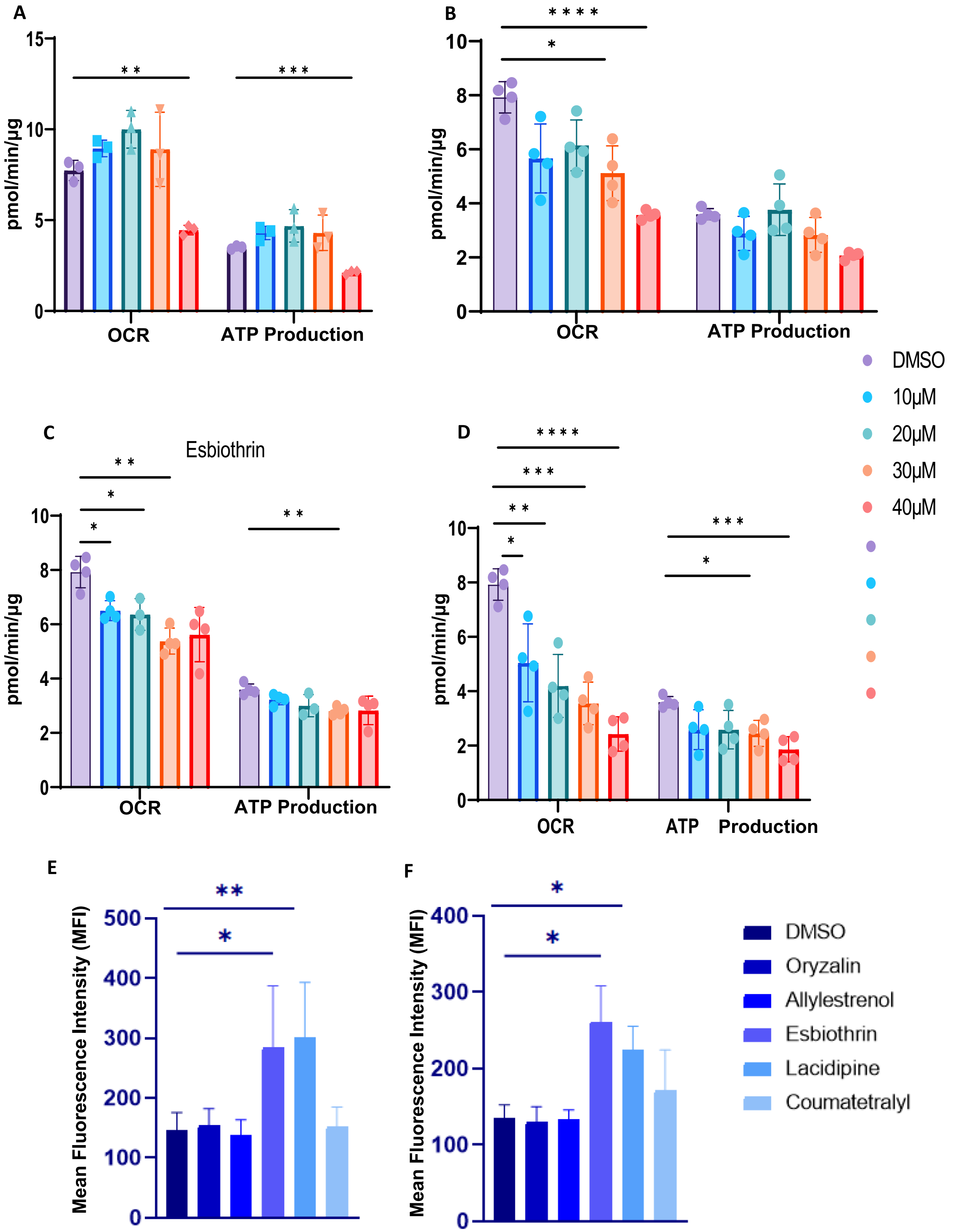


Fig. 4 (See legend on previous page.)

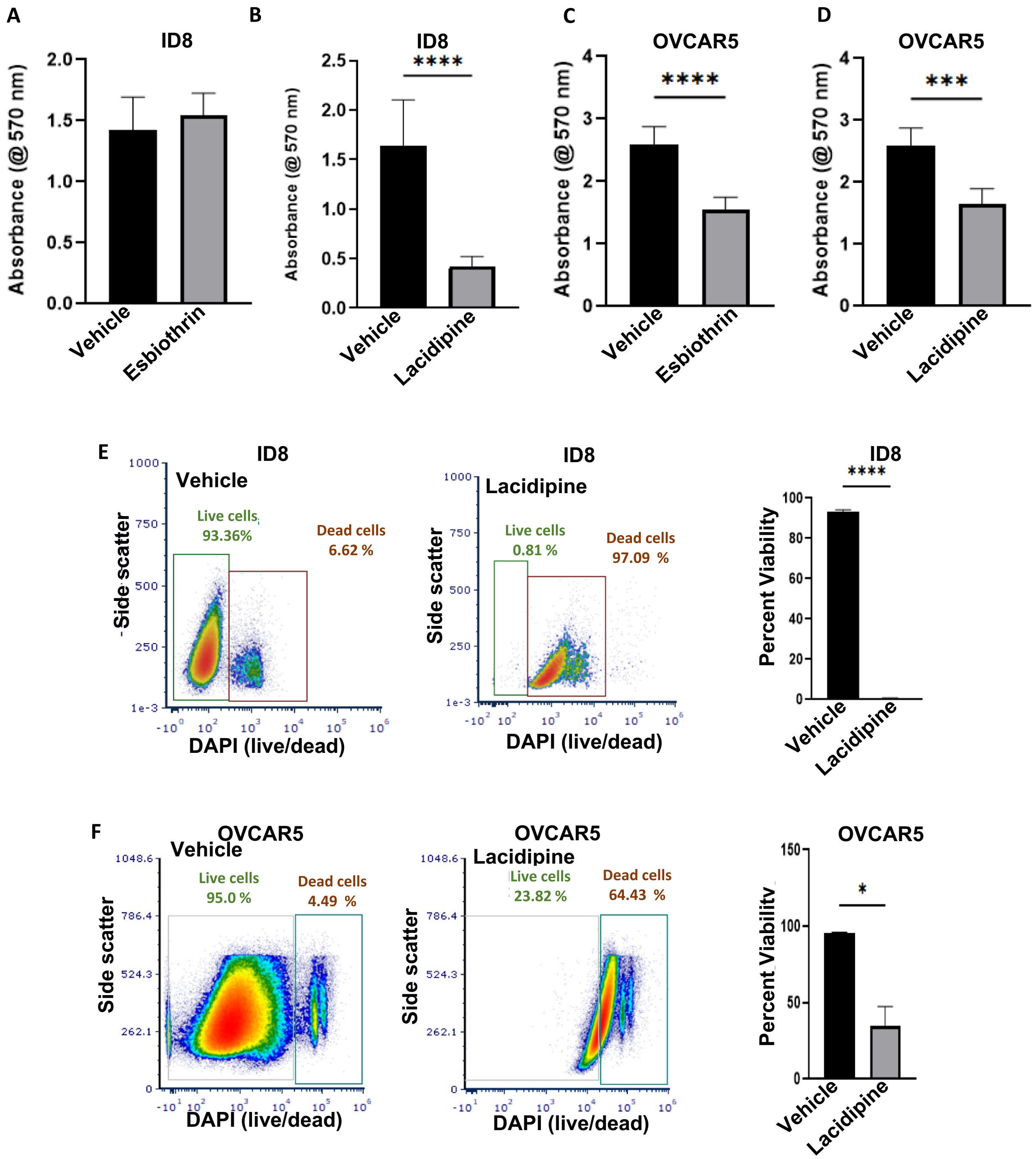


Fig. 5 Lacidipine and esbiothrin cause cancer cell death. A-D, ID8 (**A** and **B**) and OVCAR-5 (**C** and **D**) cancer cell lines were treated with esbiothrin (**A** and **C**) and lacidipine (**B** and **D**) at 10 µM (gray bars) or with vehicle (DMSO, black bars) for 72 h and proliferation of the cell lines was determined using MTT assay. The assays were conducted in triplicate and in each experiment, there were eight technical replicates of each condition tested. The plots show average data from the three biological replicates for each cell line. $p < 0.05$. E and F, ID8 (**E**) OVCAR5 (**F**) cells were treated with lacidipine (10 µM, gray bars) for 72 h and stained with DAPI to assess cell death by flow cytometry. Aggregate data from three technical replicates are shown in the bar graphs for each cell line

of its inherent fluorescence that interfered with that of H2DCFDA, oryzalin, allylestrenol, and coumatetralyl did not induce an increase in intracellular oxygen radicals.

Finally, we tested the chemotoxic effects of the six candidate compounds against OVCAR-5 and ID8 cells. In these assays, oryzalin, allylestrenol, coumatetralyl, and CCMP did not mediate statistically significant decrease in the viability of OVCAR-5 and ID8 cells after 72 h treatment (data not shown). Treatment with esbiothrin had no effect on viability of ID8 but resulted in significant decrease in viability of OVCAR-5 cells (Fig. 5A and C). On the other hand, lacidipine showed consistent and statistically significant reduction in viability of both OVCAR-5 and ID8 cells after 72 h of treatment (Fig. 5B and D).

Since consistent decrease in viability was observed with lacidipine treatment, we further tested this drug for its ability to induce cell death using a flow cytometry-based assay. In this assay, OVCAR-5 and ID8 cells were treated with lacidipine. After 72 h of incubation, the cells were labeled with DAPI, a dye that percolates through pores that form in cell membranes of dying cells. We observed that the majority of the ID8 (~97%) and OVCAR-5 (~64%) died following exposure to lacidipine (Fig. 5E and F). These experiments identified two (lacidipine and esbiothrin) of the six randomly selected compounds as potential candidates for further testing against ovarian cancer.

Pipeline-aggregated bioassay data train effective machine learning classifiers of OXPHOS inhibitors

Besides, enabling selection of candidate compounds from historical public assay data, our data mining pipeline provides rich data for training machine learning classifiers to prioritize novel compounds for testing. We observed that classifiers trained with these data perform well on predicting OXPHOS inhibitors in retrospective testing based on standard virtual screening metrics. Three standard machine learning classification models, commonly used for ligand-based virtual screening tasks, were developed using the PC_OXPHOS data subset compiled by the pipeline: logistic regression (LR), random forest (RF), and support vector machine (SVC) classifiers. For each model type, standard chemical fingerprints and/or RDKit molecular descriptors were used as compound input representations. Model hyperparameters were explored in fivefold cross-validation on 85% of the data. The top performing model from each combination of classifier type and input representation (Additional file 4) was then evaluated on a held-out test set comprising 15% of the set (Table 1). The RF and SVC models perform very well in both classification metrics F1 and MCC as well as in standard virtual screening metrics ROCAUC, average precision (like PRAUC), and EF1, which depend only on the rank order based on model scoring.

For another retrospective test of our best models, we identified known OXPHOS inhibitors from literature outside of our training set and examined our performance in virtual screening to retrieve these compounds within random sets of inactive compounds from our test set. We compiled 51 documented OXPHOS inhibitors (Additional file 5) from which 7 were removed due to overlap with the training set (boldface). Given the remaining 44 novel actives and a target baseline active fraction of 1.2% as observed in the training set, we shuffled these documented actives in with subsets of 3658 inactives randomly drawn from the test set. Both fingerprint-based RFC and SVC were evaluated on 10 sets composed with different inactive sampling (Table 2). Both models fail to retrieve any of the known positives based purely on classification label. However, most of the known actives receive high scores from the models. Thus, the models show excellent potential in virtual screening applications that rely on score-based ranking rather than discrete prediction labels. For the RFC, the ROCAUC and average precision were 0.900 (0.002) and 0.303 (0.019), respectively. The EF1 was 29.4 (4.4). As a comparison, from a random selection of 1% of compounds (37) from a set of 3702 compounds, one should expect on average 0.44 actives (fewer than 1); with hit enrichment provided by the RFC prioritization, one should expect 11–15 active compounds among the top scoring 37 (30–40%). The SVC performed worse with respect to ROCAUC and average precision, but slightly better with respect to the EF1 metric. Both RF and SVC classifiers perform worse than observed on the test set (Table 1).

Classifiers like LR, RF, and SVC are readily interpretable and thus facilitate further elucidation of the key molecular features, those most informative in discriminating OXPHOS active and inactive compounds. Feature importance analyses on our top-performing descriptor-based and fingerprint-based RF models were carried out using two methods: Mean Decrease in Impurity (MDI) and feature permutation (Tables 3 and 4). MDI (or Gini Importance) scores features based on the frequency with which a given feature appears at nodes among all of the RF trees, weighted by the number of samples split by those nodes [24]. MDI analysis of the descriptor- and fingerprint-based RF models reveals key features, which appear highly correlated with the features' absolute mean divergences between active and inactive populations (Fig. 2A and B) with respect to both descriptors (0.450) and fingerprint bits (0.908) (Additional file 6). Feature importances were also evaluated by feature permutation, a model-agnostic approach that reflects the decrease in model performance (here average precision)

Table 1 Performance of machine learning models used to predict OXPHOS inhibitors

Model	Balanced_acc	Avg_prec	ROC_auc	F1 score	MCC	EF1	NEF1
Logistic Regression							
model_l1_saga_C21.54435_desc.pkl	0.844	0.233	0.905	0.117	0.205	29.412	0.322
model_l2_saga_C1291.54967_desc.pkl	0.845	0.234	0.905	0.118	0.205	29.020	0.318
model_l1_saga_C0.04642_fps.pkl	0.847	0.335	0.906	0.172	0.256	40.000	0.438
model_l2_saga_C0.04642_fps.pkl	0.836	0.364	0.900	0.119	0.204	40.392	0.442
model_l1_saga_C0.35938_both.pkl	0.886	0.527	0.936	0.269	0.350	54.510	0.597
model_l2_saga_C0.00599_both.pkl	0.879	0.512	0.937	0.243	0.326	54.510	0.597
Random Forest							
model_RF_desc_8000_1.0_none_0.050.pkl	0.884	0.751	0.952	0.712	0.716	69.412	0.760
model_RF_fps_8000_1.0_none_0.0023.pkl	0.708	0.786	0.962	0.746	0.744	73.333	0.803
model_RF_both_8000_1.0_none_0.0010.pkl	0.686	0.782	0.956	0.755	0.754	72.157	0.790
Support Vector Machine							
model_svc_tanimoto_fps_C_10.000_gamma_null_coef0_null_degree_null.pkl	0.780	0.696	0.914	0.703	0.724	69.020	0.755
model_svc_linear_desc_C_0.001_gamma_null_coef0_null_degree_null.pkl	0.846	0.185	0.898	0.118	0.205	22.745	0.249
model_svc_linear_fps_C_0.001_gamma_null_coef0_null_degree_null.pkl	0.826	0.291	0.905	0.241	0.305	38.824	0.425
model_svc_linear_both_C_0.001_gamma_null_coef0_null_degree_null.pkl	0.842	0.393	0.914	0.302	0.358	44.314	0.485
model_svc_rbf_desc_C_10.000_gamma_scale_coef0_null_degree_null.pkl	0.846	0.739	0.936	0.745	0.746	69.804	0.764
model_svc_rbf_fps_C_15.000_gamma_scale_coef0_null_degree_null.pkl	0.819	0.752	0.927	0.751	0.761	73.725	0.807
model_svc_rbf_both_C_1.000_gamma_auto_coef0_null_degree_null.pkl	0.831	0.769	0.937	0.761	0.768	73.333	0.803
model_svc_sigmoid_desc_C_10.000_gamma_scale_coef0_-10.000_degree_null.pkl	0.498	0.132	0.887	0.022	-0.013	18.431	0.202
model_svc_sigmoid_fps_C_3.000_gamma_scale_coef0_-1.000_degree_null.pkl	0.850	0.751	0.925	0.770	0.771	72.941	0.798
model_svc_sigmoid_both_C_2.000_gamma_scale_coef0_-1.000_degree_null.pkl	0.856	0.504	0.927	0.670	0.668	63.529	0.695
model_svc_poly_desc_C_1.000_gamma_scale_coef0_0.300_degree_3.pkl	0.846	0.590	0.921	0.638	0.636	65.098	0.712
model_svc_poly_fps_C_0.100_gamma_scale_coef0_0.000_degree_2.pkl	0.846	0.748	0.954	0.755	0.755	70.980	0.777
model_svc_poly_both_C_0.100_gamma_scale_coef0_0.000_degree_3.pkl	0.827	0.728	0.956	0.746	0.750	69.412	0.760

Table 2 Virtual screening model performance on compound sets containing known OXPHOS inhibitors

Metric	RFC-fps		SVC-fps (RBC)	
	Mean	stdev	Mean	stdev
ROCAUC	0.900	0.002	0.823	0.003
Avg Prec	0.303	0.019	0.275	0.019
EF1	29.4	4.4	32.0	1.3
NEF1	0.405	0.060	0.441	0.017

when a given feature's values are randomly shuffled on a test set and then its samples re-evaluated. When ranking the features by permutation importance, we observed lower correlation with the absolute feature divergence values (0.005 sand 0.358 for descriptors and fingerprints, respectively) (Additional file 6). We take permutation as the more reliable assessment since the approach evaluates on data outside of the training set and is known to be less likely to inflate high cardinality features like MDI [25]. Nevertheless, MDI and permutation methods were correlated (0.562 and 0.351 for descriptors and fingerprints, respectively) with substantial agreement among their highest ranked features based on percentile ranks in Tables 3 and 4.

For descriptor features, quantitative estimate of drug-likeness (QED) and MolLogP are among the most important. In RF and other ensemble models, information provided by a feature is dependent on its relationship to other features and so a direct positive or negative influence on the classification score is not revealed by the importance scores in the methods used and likely oversimplifies a feature's influence in the model. However, based on the direction of shift among the strongly divergent features, it appears that QED has a profoundly negative impact on OXPHOS activity, whereas elevated MolLogP has a positive impact. Other highly ranked property-based descriptor features include several composite features involving parameters like sLogP and PEOE (partial charge contributions) with respect to van der Waals surface area (VSA) contributions—see https://greglandrum.github.io/rdkit-blog/posts/2023-04-17-what-are-the-vsa-descriptors.html for discussion on their derivations. Notably, BCUT2D_MRHI, a complex descriptor based on the largest eigenvalue for an atom adjacency matrix of empirical atom-wise molar refractivity contributions has a strong negative shift in the divergence. Interpretation of these complex descriptors is not straightforward. The extent to which a parameter like BCUT2D_MRHI relates to physical molar refractivity, which in turns relates to polarizability and molecular weight, is not clear.

Fragment-based descriptors might be more directly interpretable. Amides and generic carbonyl groups (including within carboxylate groups) are strong, likely negative, features for OXPHOS inhibition

Table 3 Feature Importances for Molecular Descriptors

Molecular descriptor	Divergence			MDI		Permutation		Percentile rank		
	μ_0	μ_1	Δ_{01}	μ	σ	μ	σ	Perm	MDI	Div
qed	0.009	− 0.738	− 0.746	0.0211	0.0262	0.0140	0.0037	1.000	0.990	0.971
VSA_EState7	− 0.006	0.508	0.514	0.0146	0.0134	0.0104	0.0012	0.995	0.961	0.821
PEOE_VSA12	0.004	− 0.320	− 0.324	0.0203	0.0284	0.0081	0.0033	0.990	0.986	0.667
MolLogP	− 0.009	0.783	0.792	0.0305	0.0318	0.0066	0.0025	0.986	1.000	0.981
SMR_VSA3	0.003	− 0.227	− 0.230	0.0173	0.0206	0.0064	0.0016	0.981	0.971	0.585
fr_amide	0.006	− 0.481	− 0.487	0.0198	0.0318	0.0059	0.0021	0.976	0.976	0.792
BCUT2D_MRHI	0.004	− 0.362	− 0.366	0.0199	0.0265	0.0047	0.0025	0.971	0.981	0.710
VSA_EState2	0.002	− 0.165	− 0.167	0.0101	0.0100	0.0045	0.0010	0.966	0.899	0.517
PEOE_VSA2	0.004	− 0.305	− 0.309	0.0077	0.0098	0.0045	0.0009	0.961	0.783	0.652
SlogP_VSA8	− 0.005	0.460	0.465	0.0094	0.0129	0.0044	0.0012	0.957	0.889	0.768
PEOE_VSA8	0.001	− 0.086	− 0.087	0.0082	0.0075	0.0043	0.0015	0.952	0.821	0.353
SlogP_VSA3	0.002	− 0.135	− 0.137	0.0091	0.0102	0.0034	0.0012	0.947	0.874	0.449
MaxPartialCharge	0.002	− 0.132	− 0.134	0.0103	0.0104	0.0034	0.0009	0.942	0.913	0.440
MinAbsPartialCharge	0.002	− 0.165	− 0.167	0.0102	0.0102	0.0034	0.0009	0.937	0.908	0.512
BalabanJ	0.001	− 0.082	− 0.083	0.0086	0.0071	0.0030	0.0009	0.932	0.841	0.333
BCUT2D_LOGPLOW	− 0.002	0.137	0.138	0.0110	0.0111	0.0029	0.0012	0.928	0.947	0.454
PEOE_VSA3	− 0.001	0.046	0.047	0.0053	0.0060	0.0026	0.0006	0.923	0.570	0.213
fr_C_O	0.003	− 0.236	− 0.239	0.0055	0.0113	0.0026	0.0003	0.918	0.604	0.599
SlogP_VSA10	− 0.002	0.133	0.134	0.0064	0.0082	0.0022	0.0009	0.913	0.676	0.444
SlogP_VSA2	0.001	− 0.058	− 0.059	0.0093	0.0088	0.0021	0.0012	0.908	0.884	0.246

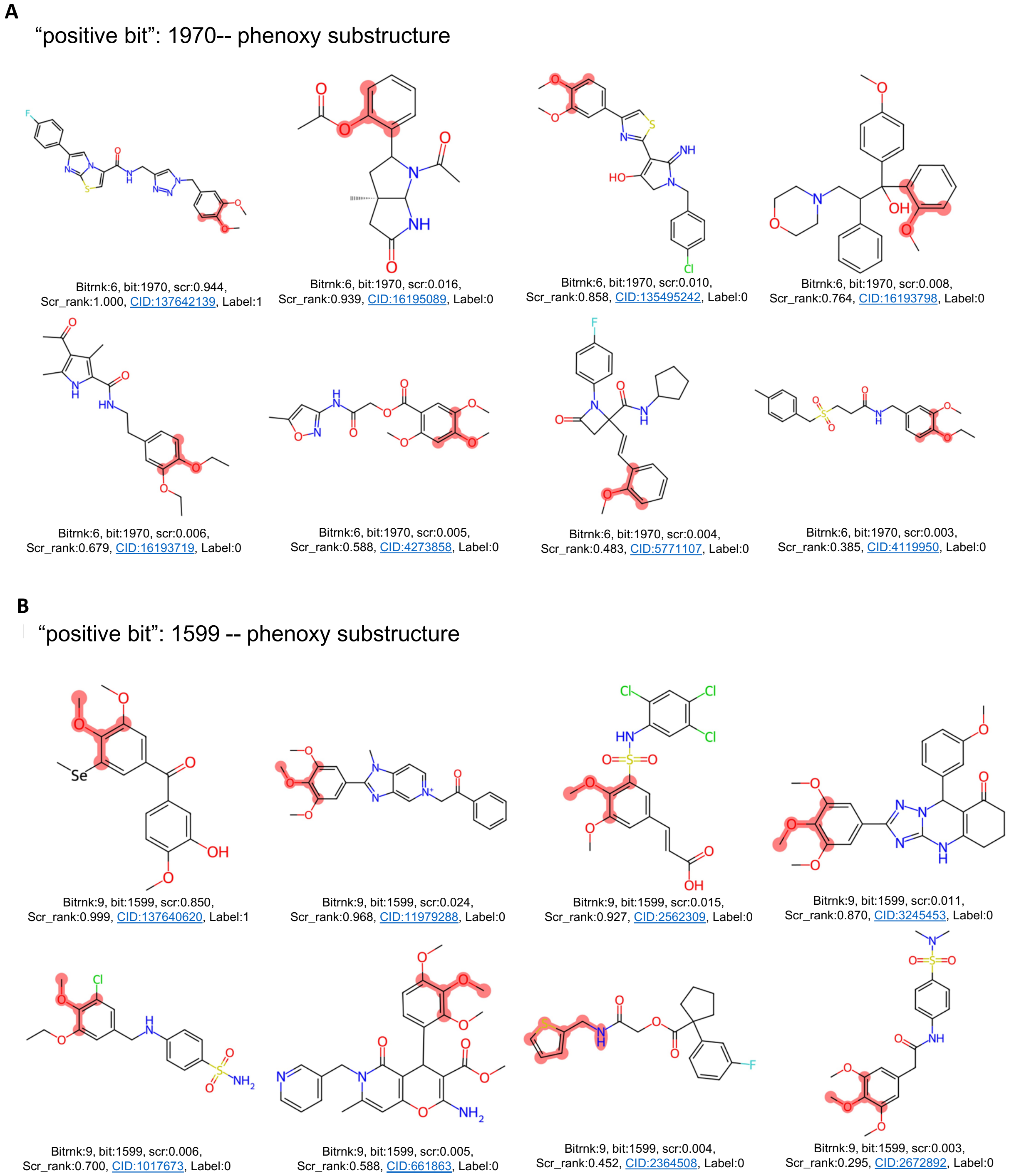
Table 4 Feature Importances for ECFP6 Fingerprint Bit Indices

Bit index	Divergence				MDI		Permutation		Percentile rank		
	μ_0	μ_1	μ	Δ_{01}	μ	σ	μ	σ	Perm	MDI	Div
1917	0.741	0.500	0.738	−0.240	0.0064	0.0102	0.00145	0.00044	1.000	1.000	1.000
378	0.512	0.470	0.511	−0.042	0.0032	0.0024	0.00128	0.00066	1.000	0.994	0.940
807	0.812	0.662	0.810	−0.150	0.0037	0.0050	0.00124	0.00055	0.999	0.998	0.997
1152	0.614	0.368	0.611	−0.246	0.0067	0.0099	0.00124	0.00107	0.999	1.000	1.000
935	0.602	0.496	0.601	−0.106	0.0036	0.0031	0.00118	0.00037	0.998	0.997	0.991
1816	0.293	0.136	0.291	−0.157	0.0041	0.0063	0.00114	0.00045	0.998	0.999	0.998
1970	0.166	0.277	0.167	0.111	0.0021	0.0025	0.00114	0.00046	0.997	0.984	0.992
651	0.008	0.030	0.008	0.022	0.0007	0.0015	0.00101	0.00004	0.997	0.808	0.836
1480	0.287	0.139	0.285	−0.148	0.0036	0.0052	0.00097	0.00059	0.996	0.996	0.997
1599	0.036	0.124	0.037	0.088	0.0022	0.0045	0.00094	0.00010	0.996	0.985	0.988
41	0.222	0.048	0.220	−0.173	0.0049	0.0095	0.00088	0.00018	0.995	0.999	0.999
25	0.048	0.104	0.049	0.056	0.0014	0.0021	0.00081	0.00025	0.995	0.951	0.967
926	0.504	0.410	0.503	−0.095	0.0031	0.0025	0.00077	0.00053	0.994	0.993	0.990
1296	0.014	0.076	0.014	0.063	0.0012	0.0031	0.00059	0.00002	0.994	0.940	0.974
604	0.009	0.048	0.009	0.040	0.0010	0.0024	0.00057	0.00008	0.993	0.899	0.934
1765	0.089	0.141	0.090	0.052	0.0015	0.0017	0.00056	0.00011	0.993	0.957	0.961
961	0.047	0.118	0.048	0.071	0.0016	0.0026	0.00054	0.00026	0.992	0.969	0.980
694	0.123	0.250	0.124	0.127	0.0029	0.0042	0.00054	0.00035	0.992	0.993	0.994
1722	0.356	0.221	0.354	−0.135	0.0034	0.0039	0.00053	0.00044	0.991	0.996	0.995
1764	0.018	0.046	0.019	0.027	0.0008	0.0014	0.00052	0.00012	0.991	0.845	0.876

activity (ranking 6th and 18th overall based on permutation analysis). However, more specific carbonyl types represent the strongest positive fragment descriptors: allylic oxide (fr_allylic_oxid) and ketones (ranking 40th and 50th). Important ECFP6 fingerprint bits/substructures, when mapped visually onto compound structures, are congruent with our calculated descriptor fragment importances. Based on the permutation importance rank order of bits in ECFP6 fingerprints, phenoxy substructures were mapped to positive bits 1970 and 1599 (Fig. 6A and B). Another high-importance positive bit, 694, maps to a general methine carbon (e.g., alkenes, carbonyl, and imine bonds) (Additional file 7A). Prominent negative bits, 1917 and 1152 map to carbonyl and secondary amine substructures (Additional file 7B and C), respectively. Note, some collisions are observed at certain bits, reflected by different substructures mapping to the same bit index. However, we observe this to be infrequent and should not severely complicate interpretation of bit importance, given our fingerprint length (2048) and resulting bit density.

Discussion

Here, we report the successful development of a pipeline for data mining of public assay records in PubChem to identify diverse small molecule chemotypes with high OXPHOS inhibitory potential. The data compiled by the pipeline were also useful for training LBVS models for predicting OXPHOS inhibition. Our in-silico approach requires practically no new resource investment in assays to cull a total of 312,039 CIDs to a final hit list of 1859 CIDS (1852 unique parent SMILES) that not only have high potential for activity as OXPHOS inhibitors but also pass PAINS filters and possess generally good PK properties for oral bioavailability. Biological testing of six compounds selected from distinct active chemotypes shows a high hit rate as (a) four of these compounds inhibited OXPHOS (via measurements on OCR and ATP levels) and (b) two of the four compounds that inhibited OXPHOS also elevated intracellular oxygen radicals and decreased the viability of ovarian cancer cell lines. With the majority of the 1859 compounds identified through our data-mining pipeline still to be tested, it is likely that several



“positive bit”: 1970-- phenoxy substructure

Bitrnk:6, bit:1970, scr:0.944, Scr_rank:1.000, CID:137642139, Label:1

Bitrnk:6, bit:1970, scr:0.016, Scr_rank:0.939, CID:16195089, Label:0

Bitrnk:6, bit:1970, scr:0.010, Scr_rank:0.858, CID:135495242, Label:0

Bitrnk:6, bit:1970, scr:0.008, Scr_rank:0.764, CID:16193798, Label:0

Bitrnk:6, bit:1970, scr:0.006, Scr_rank:0.679, CID:16193719, Label:0

Bitrnk:6, bit:1970, scr:0.005, Scr_rank:0.588, CID:4273858, Label:0

Bitrnk:6, bit:1970, scr:0.004, Scr_rank:0.483, CID:5771107, Label:0

Bitrnk:6, bit:1970, scr:0.003, Scr_rank:0.385, CID:4119950, Label:0

“positive bit”: 1599 -- phenoxy substructure

Bitrnk:9, bit:1599, scr:0.850, Scr_rank:0.999, CID:137640620, Label:1

Bitrnk:9, bit:1599, scr:0.024, Scr_rank:0.968, CID:11979288, Label:0

Bitrnk:9, bit:1599, scr:0.015, Scr_rank:0.927, CID:2562309, Label:0

Bitrnk:9, bit:1599, scr:0.011, Scr_rank:0.870, CID:3245453, Label:0

Bitrnk:9, bit:1599, scr:0.006, Scr_rank:0.700, CID:1017673, Label:0

Bitrnk:9, bit:1599, scr:0.005, Scr_rank:0.588, CID:661863, Label:0

Bitrnk:9, bit:1599, scr:0.004, Scr_rank:0.452, CID:2364508, Label:0

Bitrnk:9, bit:1599, scr:0.003, Scr_rank:0.295, CID:2672892, Label:0

Fig. 6 Important ECFP6 bits **A** 1970 and **B** 1599, corresponding to phenoxy-type substituents, are highlighted on 8 compounds drawn from the PC_OXPHOS set. The 8 example compounds representing the given bit were intentionally selected to span the range of scoring by our RFC model. Bits 1970 and 1599 were particularly abundant in OXPHOS-active compounds. The red highlighted atoms and bonds correspond to the substructures represented by each bit. Under each compound several parameters are reported. *Bitrnk* is the ranked importance of the bit according to permutation analysis. *Bit* is the bit index within the ECFP6 fingerprint (length = 2048). *Scr* is the score observed for the compound by our top performing RFC. *Scr_rank* is the percentile ranking of the score for the compound. *CID* is the PubChem Compound Identifier. *Label* is the ground truth label for the compound (1 = active, 0 = inactive)

additional OXPHOS inhibitory candidates can be identified as potential therapeutic candidates for ovarian and other cancers.

Furthermore, the pipeline revealed distinctive chemical features of OXPHOS inhibitors, provided a breakdown of diverse OXPHOS-active chemotypes (464 clusters) observed across 8415 public assay records (Additional file 2), mapped these to specific OXPHOS-active high-density regions within the broader chemical testing space, and provided a rich data for training classifiers that performed well in retrospective tests and should have future applicability in prioritizing OXPHOS inhibitor candidates within new compound sets.

Observations on OXPHOS inhibitory chemotypes and their chemical characteristics

Our data mining pipeline retrieved a substantial fraction of compounds from a set of documented OXPHOS inhibitors we compiled from literature. Among the 51 inhibitors (Additional file 5), 20 of these were in the full compound set (312,039 unique CIDs) based on canonical SMILES matches. Of these, 10 were appropriately recognized as active based on our pipeline conditions and assigned positive labels: atovaquone, plumbagin, FCCP, alpha-TOS, antimycin A, rotenone, curcumin, deguelin, Atpenin A5, and fenofibrate. However, our activity labeling conditions also resulted in 3 documented inhibitors being mislabeled as "inactive:" phenformin, citral, and 3-nitropropionic acid. Furthermore, several known actives had mixed results in the compiled assay records and were considered "unlabeled." These included metformin, diphenylamine HCl, lonidamine, amobarbital, tiabendazole, papaverine, and ranolazine. For purposes of compound prioritization and training machine learning models, we favored stringent labeling assignment conditions so that positive and negative instances were unambiguous with respect to assay outcomes in their associated AIDs. However, the cause of these apparent mislabels is not clear, potentially resulting from assay sensitivity or error, inclusion of non-relevant assays, excessive stringency of labeling criteria, or other issues. Nevertheless, given the high false positive/negative rates in primary screening data, and the large set of disparate assays from which we collected data, we are encouraged that 10 of the 20 documented inhibitors in the mined data set were assigned active labels and another 7 were considered inconclusive. Note that active labeled compounds represented only 1.5% of the initial data set (4794 actives in 312,039 total CIDs).

Among the 20 documented inhibitors identified in the original data set, all had at least one ETC-relevant assay record. Interestingly, 11 of these 20 fail the PK filters and 3 fail the PAINS filter: atovaquone (adjacent carbonyls), plumbagin (quinone), and FCCP (imine and cyano groups). Of the 20, only 10 were assigned active labels based on activities across compiled assay records, and just 5 were ultimately included in the final 1859 actives after passing PAINS, PK, and ETC-relevance conditions: rotenone, curcumin, deguelin, atpenin A5, and fenofibrate. Rotenone and deguelin are multicyclic structures with α, β-unsaturated carbonyls. Curcumin, and Atpenin A5 are monocycles with a α,β-unsaturated carbonyls. Fenofibrate has a ketone group bridging 2 different phenyl substituents. Rotenone, deguelin, and fenofibrate target Complex I in ETC and have documented antiproliferative effects whereas Atpenin A5 targets Complex II and exhibits cardioprotective effects (reviewed in [13, 26–29]).

Compounds with mitochondrial bioavailability and OXPHOS activity tend to possess distinct physicochemical features. A previous study involving models for predicting mitochondrial toxicity showed significant enrichment for mitochondrial toxins in the high SlogP range of 4–9 [30]. Toxins also had a higher abundance of aromatic rings/aromatic carbocycles, higher molecular weight and surface areas (LabuteASA), and lower fraction sp^3 hybridized carbons. Our analysis confirms most of these earlier findings, further supporting unique property ranges for compounds intended for distribution to the inner membrane of mitochondria. In terms of RDKit physicochemical descriptors, OXPHOS actives exhibit relatively high MolLogP, RingCount, Labute ASA, MW, and NumAromaticCarbocycles, (Fig. 2A). NumAromaticCarbocycles and MW are not shown but were shifted right by 0.357 and 0.580 standard deviations, respectively. Contrary to a previous report where mitotoxins were shown to possess a lower fraction of Csp3 carbons, we observed a very small increase in the fraction Csp3 (0.059) in OXPHOS actives [27]. Fragment analysis revealed enrichment for bicyclic, phenol, Ar-OH, and benzene substructures (Fig. 2B). Whereas previous 2-D PCA projections did not show notable separation of mitotoxins from non-toxins [30], our analysis via 2-D UMAP on a much larger set of compounds did show clear distinctions in OXPHOS-active and -inactive distributions within chemical space (Fig. 3).

Observations on testing of 6 prospective OXPHOS inhibitors

The six compounds that we tested represent distinct chemotypes. While four of these six compounds showed OXPHOS inhibitory activity, it is important to note that one of the compounds from this group, esbiothrin, is a known pesticide and hence is likely to present an unacceptable toxicity profile in human patients. This situation can be conveniently remedied in future iterations of our data mining algorithm by including additional filters that specifically exclude compounds that are known to be pesticides, insecticides, or those with a reported high toxicity in humans or in experimental animals.

Among the six compounds we have tested, lacidipine is especially promising as it exhibits higher activity as an OXPHOS inhibitor and shows clear chemotoxic activity against cancer cells. Extensive studies have demonstrated lacidipine as an antagonist of the L-type calcium ion channels and because of this function, lacidipine is used for the treatment of hypertension. Our data mining approach and the data from the biological assays demonstrate for the first time a novel biological activity for lacidipine.

Our data mining pipeline reveals that OXPHOS inhibitory chemotypes are diverse (represented by 464 total structural clusters (Additional file 3A) and 135 larger clusters (greater than 2 members) (Additional file 3B). However, our UMAP projection (Fig. 3) shows that these clusters tend to localize to specific, confined regions of the chemical testing space in superclusters. Inspection of the compound structures within these hot spots confirms the chemotypes are highly related within these regions. This mapping and confinement of the OXPHOS-active space might be leveraged for focusing screening efforts for other OXPHOS inhibitors. Such mappings may also serve as a basis to more deeply characterize the OXPHOS structure–activity relationship (SAR) of active chemotypes—perhaps determination of their specific biochemical targets within the OXPHOS pathway. Importantly, this analysis also identifies the chemical space that is productive when targeting OXPHOS pathway.

Limitations in data mining prototype and planned advancements

Given our ongoing effort to develop drug candidates for treatment of ovarian cancer, we believe the pipeline might be improved by addressing some limitations exposed in prototyping here. In compound prioritization, greater emphasis should be placed on achieving PK/tox objectives, in addition to OXPHOS inhibitory potential, to facilitate faster translation to in vivo testing. Perhaps, priority should be given to compounds from OXPHOS-active chemotypes represented by already approved drugs or known metabolites. Also, many of the OXPHOS active agents collected in data mining were later observed to have known liabilities with respect to toxicity. This might be something gleaned directly from CID records, which are frequently annotated with toxicity data. Additionally, more stringency might be beneficial in AIDs by requiring cancer-related terms like anti-cancer, anti-proliferative, anti-tumor, cell killing, etc. Further, since OXPHOS, mitochondrial function, and oxidative stress are terms likely associated with toxicity- or counter-screens, an effort should be made to flag AIDs involving toxicity screens. For example, AIDs involving healthy tissues/normal cells (for e.g., fibroblasts or PBMCs), or in animal models and cell death read-outs based on dye reagents for assessing cell penetrance or DNA binding [31], should be flagged as such. Inclusion of an additional toxicity label may enable multi-objective prioritization based on both potential for OXPHOS activity and potential toxicity.

Another key issue we hope to address stems from the fact that the OXPHOS inhibitors identified in our data mining pipeline act at different points in the electron transport chain, or perhaps even outside of the chain. In fact, some of the agents have been characterized with respect to their targeted complexes within OXPHOS [13]. The desired oxidative stress condition or metabolic perturbation, and thus the anti-cancer effect ultimately promoted by the agents, likely depends on their specific biochemical targets in the chain. Therefore, we hope to advance this work by elucidating the SAR between chemotypes and points of intervention in the OXPHOS pathway. Such efforts are likely critical for improving prioritization of candidate compounds with the highest potential as starting points for chemotherapeutic agents for the treatment of ovarian and other cancers.

Conclusions

We have successfully developed a data mining and machine learning-based approach for efficient identification of OXPHOS inhibitors. In a data set mined from PubChem, we observe a broad range of potentially OXPHOS inhibitory chemotypes, exhibiting signature chemical substructures and properties. Using UMAP with fingerprint representations, these active chemotypes concentrate in specific regions of chemical space. Testing of six compounds taken from distinct active regions in the OXPHOS-active UMAP space resulted in 2 compounds that promote oxidate stress conditions, one of which with anti-proliferative activity in ovarian cancer cell lines. Here we demonstrate that our data mining approach enables highly focused screening for OXPHOS inhibitors with resulting anti-cancer activities. This prototype pipeline is extensible and could be adapted for focus screening on other phenotypic targets for which sufficient public data are available.

Methods

Extraction and processing of PubChem bioassay data

The complete bioassay data extraction and mining workflow to identify OXPHOS inhibitor chemotypes is shown in Fig. 1B. The pipeline was implemented on an Ubuntu 20.04.6 LTS workstation through a sequence of linux bash shell and python scripts (Python 3.8.10). The conda environment (.yaml), scripts, and documentation are available on GitHub: https://github.com/SpencerEricksen/PubChem_DataMining_OXPHOS. The main modules (versions) used were RDKit (v2022.09.1), pandas (v1.5.3), and scikit-learn (v1.2.2). The pipeline was initiated by submitting the following query to PubChem via the browser-based advanced assay search (https://www.ncbi.nlm.nih.gov/pcassay/advanced) on Feb 22, 2022, 13:57 CST):

> "electron transport chain"[Assay Description] OR
> "mitochondrial complex"[Assay Description] OR
> "mitochondrial respiratory chain"[Assay Description] OR
> "mitochondrial membrane potential"[Assay Description]) AND
> small_molecule[filt]

The query returned a list of matching Assay IDentifier (AID) records having the OXPHOS-related search terms in their bioassay description text. The number of matching AIDs (8415) exceeded the number of full AID records permitted for download in a single request (1000) so the matching AID information was saved using the "Send to" option to a text file ("pcassay_result.txt"). This text file includes a list of AIDs with associated assay titles, data source, number of substance records (SIDs) tested, and the number of active SIDs in the assay. Using a list of the AIDs from this file, the complete assay records–including compound records tested (CIDs) with associated assay outcomes–were downloaded (.csv) using a python wrapper to initiate URL-based requests through PUG-REST for each AID in the list.

The AID data sets were then merged into a single dataframe (bioassay dataframe) where each row represents a single compound testing record within a specific assay. Then, an xml-style SMILES pc_fetch file (.cgi) was created for the union of CIDs in the dataframe and submitted via PUG REST to fetch the SMILES. Returned SMILES were then merged into the bioassay dataframe using the CID as they key. Using the RDKit python cheminformatics toolkit, the PubChem SMILES were canonicalized, sanitized, and stripped of smaller counter ions. Some counter ions were not included among standard salt counter ion definitions in RDKit and so a secondary salt stripper function was applied that retained only the largest molecular fragment in cases where multiple fragments remain present in a SMILES. The RDKit-canonicalized SMILES, both salt and desalted forms, were stored in the bioassay dataframe as additional columns.

Next, bioassay descriptions and metadata associated with each AID were downloaded from PubChem in.xml format using curl within a BASH wrapper. From each AID's.xml file, the following bioassay fields were extracted and merged into the bioassay dataframe using the AID as the key: 'id', 'pmid' (PubMed Identifier), 'DOI' (Digital Object Identifier), 'Year', 'ChEMBL Target Name', 'ChEMBL Target Type', 'Target ChEMBL ID', 'Confidence', 'Relationship Type', 'Title', 'name', and 'Abstract'.

To prioritize the most relevant assays, AIDs that specifically involved electron transport chain (ETC) protein targets or ETC-associated functions were flagged. ETC-linked AIDs were identified by searching text from each AID's name, title, and abstract for matches to a list of 39 positive ETC-related terms and 2 negative, photosynthesis-related terms (Additional file 1). A list of PubMed IDs (PMIDs) associated with these ETC-linked AIDs was generated. The papers (PMIDs) were examined for experimental relevance to ETC. Where the papers were confirmed to be ETC-relevant, the AID was given highest priority and flagged as having an ETC-linked PMID. For purposes of data mining and classifier training, only data from ETC-linked AIDs were used but an ETC-linked PMID was not required.

The set of CIDs in the bioassay dataframe and their corresponding desalted, RDKit-canonicalized SMILES were then isolated for generation of structural properties and features. Using primarily RDKit methods, each compound was assigned its Bemis-Murcko and reduced Murcko scaffolds, natural product-likeness score [32], PAINS flags using RDKit substructure matching against list of PAINS substructures (https://github.com/iwatobipen/rdkit_pains [33, 34], RDKit molecular descriptors (Additional file 8A), Morgan (circular) chemical fingerprints of length 2048 and radius = 3 [35], and a conditional Lipinski-like PK property flag based on desirable physicochemical constraints typical in hit-to-lead workflows [36]: molecular weight > 200 Da, mLogP < 5.8, total polar surface area < 150 Å^2, fewer than 20 heavy atoms (non-hydrogen), at least one nitrogen or oxygen atom, and at least 1 ring.

Next, activity outcomes and scores within the 8415 OXPHOS-related AIDs were compiled and associated for each unique CID. Based on the compiled activity data, an activity label was designated for each CID. CIDs with at least 1 'Active' outcome and a number of active outcomes greater than or equal to the number of inactive outcomes were considered OXPHOS-active molecules (inhibitors). Molecules with at least one inactive outcome and no active, unspecified, or inconclusive outcomes were designated as OXPHOS-inactive. These labels were used for data mining and machine learning procedures.

Clustering of OXPHOS-active compounds

To examine the chemotypes active specifically in OXPHOS inhibitory function, the set of compounds (unique PUBCHEM_CIDs) labeled as OXPHOS-active, derived from ETC-linked AIDs, and passing PAINS and our PK filters were clustered by structural similarity. Hierarchical agglomerative clustering was performed using SciPy's (v1.5.3) average linkage implementation. A condensed distance matrix was obtained based on chemical fingerprints as compound representations. As input to clustering, a pairwise distance matrix was computed for these compounds using the Jaccard distance between fingerprints (2048-length bit vectors). Medoid molecules for each cluster were selected as the compound having the lowest sum of distances to all cluster cohortss. The distance cutoff used for cluster assignment was 0.750. That means clusters with centroids nearer than 0.750 in Jaccard distance were merged into the same cluster. Based on the clustering outcome, the cluster IDs and cluster medoid flags were merged with the existing compound data.

Development and evaluation of machine learning classifiers

LR, RFC, and SVC models were developed with training data produced by the data mining pipeline. Starting with the complete set of unique compounds from the data mining pipeline (312,039 compounds), a subset of 155,653 unique compounds were isolated (PC_OXPHOS), having definitive positive (1859) or negative (153,794) activity labels ('label' = 0/1), satisfying PK and PAINS conditions, and associations with at least one ETC-linked AID.

Molecules in this set were featurized as RDKit ECFP6 (length 2048, radius = 3) fingerprints and RDKit molecular descriptors. The "Ipc" descriptor from RDKit was removed due to scaling issues. Standard scalers were fit to training data to transform raw feature values into z-scores (mean = 0, standard deviation = 1). The standard scalers were saved and applied later to features in test and calibration sets prior to input for model evaluation.

The data were then split with 85% (132,305 total, 1604/130701 active/inactive) reserved as a training set for model hyperparameter determinations and 15% diverted as a test set for subsequent classifier evaluation. Grid-based hyperparameter searches were performed for each model type as indicated in Additional file 4. Models were generated using fingerprint inputs, descriptor inputs, and both. All models used SciKit-Learn implementations save for a bespoke Tanimoto kernel (https://github.com/gmum/pykernels/blob/master/pykernels/regular.py) used in a fingerprint-based SVC model (no standardization). Best performing hyperparameter sets were selected based on the sum of ROCAUC and Average Precision summary metrics observed in fivefold CV on the training data. Models were evaluated on the test set using standard classification performance metrics, F1score (F1) and Matthew's Correlation Coefficient (MCC), as well as common ranking-based, virtual screening metrics: average precision (AP), area under the Receiver-Operator Characteristic Curve (ROCAUC), enrichment factor 1% (EF1), and normalized enrichment factor 1% (NEF1) [37].

Due to the large class imbalance in the data set (1.2% active), the RFC models required calibration for proper assessment of threshold-dependent metrics like F1 and MCC. LRs and SVCs did not require calibration. To calibrate RFCs, the data were split 3 ways into a training (70%), calibration (15%), and test (15%) sets. RFCs were trained on the smaller training set (70% vs. 85%) and then calibrated using SciKit-Learn's CalibratedClassifierCV class ('prefit' = True) on the calibration data set. Calibrated RFCs were then applied on the test set to evaluate F1 and MCC performance metrics.

UMAP analysis

To visualize and compare active and inactive compound distributions in chemical space, ECFP6 fingerprints from the same molecule set used for ML testing and evaluation were projected into 2 dimensions using the umap-learn python implementation of Uniform Manifold Approximation and Project (UMAP) analysis [38]. A grid-based search for suitable *n_neighbors* and *minimum distance* parameters was performed over the following space: {*n_neighbors*: 5, 10, 15, 20, 25, 30, 40, 50, 100, 200} and {*minimum_distance*: 0.001, 0.025, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50}. Based on inspection of the resulting maps and co-localization of cluster cohorts, suitable parameters were found to be *n_neighbors* = 50 and *minimum_distance* = 0.25. For contour maps, kernel density estimates for each compound set (all, inactives, actives) were calculated using the gaussian_kde function in Scipy's stats module.

Cell lines and reagents

The human ovarian cancer cell line, OVCAR-5, was obtained from ATCC, along with ID-8 cells (mouse ovarian cancer cell lines). These established cell lines were maintained in their respective recommended media and maintained at 37 °C in a humidified atmosphere with 5% CO_2. All cell lines of human origin were authenticated using STR (Single Tandem Repeats) analysis conducted in-house through the Translational Science Biocore shared resource of the University of Wisconsin Carbone Cancer Center. Mycoplasma testing of all cell lines was performed routinely and whenever cultures of new cell stocks were initiated.

Cell culture media and general laboratory supplies and reagents were purchased from ThermoFisher (Waltham MA) unless otherwise stated. Reagents for the Seahorse experiments were obtained from Agilent Technologies. The vendors for the six test compounds were as follows—oryzalin (Aldrich CPR, St.Louis, MO), allylestrenol (TargetMol Chemicals Inc.,Wellesley Hills, MA), esbiothrin/ allethrin and coumatetralyl (Sigma-Aldrich, St.Louis, MO), lacidipine (Ambeed, Inc, Arlington Heights, IL), and 2-cyano-*N*-cyclohexyl-3-[5-(4-morpholinyl)-2-furyl]-2-propenamide (CCMP) (Vitas-M Laboratories Ltd, Champaign, IL). All test compounds were purchased through the Aldrich Market Service (Small Molecule Screening Library Service) and except for oryzalin where data was not provided by the vendor, the purity of each compound was >92%.

Bioenergetics assays

Oxygen Consumption Rate (OCR): ID-8 (5000/well) were seeded in XFe96 Seahorse Analyzer Culture Plate overnight and next day treated with drugs (oryzalin, allylestrenol, esbiothrin/ allethrin, lacidipine, coumatetralyl, (Z)-2-cyano-*N*-cyclohexyl-3-[5-(4-morpholinyl)-2-furyl]-2-propenamide) (CCMP) at concentrations between 10–40 µM for 45 min. After treatment, cells were washed twice with XF RPMI media containing 1 mM pyruvate, 2 mM glutamine and 10 mM glucose (Agilent Technologies). Culture media was replaced with 180 µl of XF RPMI media and equilibrated for 45 min at 37 °C in a non-CO_2 incubator. The mitochondrial stress test was performed with injections of oligomycin (1 µM); FCCP (2 and 0.5 µM); and rot/antimycin (1 µM). Oxygen Consumption Rate (OCR) and ATP production were measured and analyzed with Agilent Wave software. The data were normalized to the total protein (µg) in each well.

Intracellular oxygen radical quantitation

Reactive Oxygen Species (ROS) in the mouse ID8 and human ovarian cancer cell lines OVCAR 5 was determined using the oxygen radical-sensing dye 2,7-Dichlorodihydrofluorescein diacetate (H_2DCFDA) (Sigma Aldrich) by Flow cytometry. Cells (5×10^5) were washed twice with 1X HBSS, followed by incubation with 10 µM H_2DCFDA dye in serum free media for 30 min at 37 °C under 5% CO_2. After incubation, cells were washed twice with 1X HBSS and protected from light. Test compounds at desired concentrations were added to the cells in serum-free media and after 45 min incubation, the cells were analyzed on an Attune (ThermoFisher) spectral flow cytometer. Data was analyzed using FCS Express 7 software.

Cell viability assays

Cell viability was assessed using the MTT assay as described in our previous studies. OVCAR-5 (2500 cells), ID-8 (1000 cells) were seeded in a flat bottomed—96 well plates and incubated overnight at 37 °C in 5% CO_2. Control cells were exposed to the relevant concentration of DMSO (vehicle) and test cells were treated with the respective concentration of each of the test compounds. The treatment was performed for 72 h. Following incubation, the MTT reagent (0.5 mg/ml) was added to each well and incubated for 3 h at 37 °C. DMSO (100 µl) was added to each well to dissolve the formazan crystals. Plate was mixed gently for 4–5 min and the optical density of solution in each well was determined using a microplate reader at 560 nm.

Cell viability was also tested by flow cytometry. Cell lines treated with the test drugs were harvested from the culture plates and the dead cell indicator DAPI (2 µl of 5 µg/ml stock in FACS buffer) was added to the cell suspension (100 µl). The cells were analyzed by flow cytometry on the Attune (ThermoFisher) cytometry and the data was analyzed using FCS Express 7 software.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00906-0.

Additional file 1. ETC-linkage terms applied to AID "name", "title", or "abstract". List of terms for a secondary filter consisting of 39 positive terms (involving electron transport in mitochondria) and 2 negative terms (related to the photosynthesis pathway).

Additional file 2. Structures of medoids of each compound cluster. Structures of each of the medoid that are the central compounds of each cluster identified to contain potential OXPHOS inhibitors.

Additional file 3. **A**–**F** UMAP projections of OXPHOS inhibitor candidates. Different 2D UMAP projections of the 2-D chemical space based on the PC_OXPHOS compound set. A-C show the full, inactive, and active compound distributions as KDE represented by contour maps, respectively. In A, points are overlayed for active compounds. Colored points represent different clusters exceeding 2 molecules. Remaining active molecules are shown as grey points. Projections D-F show difference maps derived by subtracting the KDEs of the active and inactive points. Positive values (red) show regions of relatively high OXPHOS-active density. Negative values (blue) show relatively low density for OXPHOS actives. Medoids from active chemical clusters of size 3 or greater are overlayed as labeled points. In D, the OXPHOS-active oxygen-rich saturated polycycles are highlighted. In E, active nitrogenous heterocycles are labeled. Map F shows some miscellaneous OXPHOS actives that do not fit into the latter categories.

Additional file 4. Validation of top performing machine learning models. Model hyperparameters were explored in 5-fold cross validation on 85% of the data and the top performing model from each combination of classifier type and input representation was evaluated.

Additional file 5. Compounds used in training set. List of 51 documented OXPHOS inhibitors and how they score relative to the test set in both RFC and SVC models. 44 of these documented inhibitors were not in the original training set and therefore serve as an additional test set for the models. True labels and classification labels are also provided in addition to model probability scores.

Additional file 6. Feature importance correlations between methods. MDI analysis of the descriptor- and fingerprint-based RF models reveals key features, which appear highly correlated with the features' absolute mean divergences between active and inactive populations with respect to descriptors and fingerprint bits.

Additional file 7. **A**–**C**. Important ECFP6 bits based on feature permutation analysis of chemical fingerprints in our RFC model. Important ECFP6 bits based on feature permutation analysis of chemical fingerprints in our RFC model. Bits (A) 694, (B) 1917, and (C) 1152 correspond to methine carbons, generic carbonyls, and generic secondary amines, respectively. The red highlighted atoms and bonds correspond to the substructures represented by each bit. Under each compound several parameters are reported. *Bitrnk* is the ranked importance of the bit according to permutation analysis. *Bit* is the bit index within the ECFP6 fingerprint (length=2048). *Scr* is the score observed for the compound by our top performing RFC. *Scr_rank* is the percentile ranking of the score for the compound. *CID* is the PubChem Compound Identifier. *Label* is the ground truth label for the compound (1=active, 0=inactive).

Additional file 8. RDKit Molecular Descriptor Definitions. Descriptors are defined for property-type and fragment-type features on separate sheets. Where available, a reference is provided for the property-type feature origins.

Acknowledgements

The authors thank the University of Wisconsin Carbone Cancer Center Small Molecule Screening Facility (Drug Development Core) and the Flow Cytometry shared resources for use of their facilities and services and for the expertise provided. We also acknowledge the computational resources provided by the University of Wisconsin-Madison's Center for High-Throughput Computing (CHTC).

Author contributions

SSE contributed to motivation for study, design of data mining procedure, data mining and machine learning codes, and data analysis; SS contributed to design and execution of biological experiments, and data analysis. LF and AK contributed to data mining and inspection of literature. NB assisted with bioenergetics experiments. LB provided input on interpretation of biological data. MSP contributed to the motivation for study, design of biological experiments, and data analysis. All authors contributed to writing the manuscript and have given approval to the final version of the manuscript.

Funding

This research was supported by a VA grant I01BX005627 to MSP and SSE and by R01GM135631 (to Dr. Anthony Gitter, not listed as co-author) and 1R01CA238423 to LB from the National Institutes of Health. The studies were also supported by generous support from the Diane Lindstrom Funds (001), Wisconsin Ovarian Cancer Alliance (001) and the Department of Obstetrics and Gynecology, University of Wisconsin-Madison. Funds from P30 CA014520 supported the shared services provided by the Drug Development Core (Small Molecule Screening Facility) and Flow Cytometry Lab of the University of Wisconsin-Madison Carbone Cancer Center.

Availability of data and materials

The original data downloaded from PubChem and ultimate pipeline-processed datasets are available on Zenodo (https://doi.org/10.5281/zenodo.11003006). Python scripts with documentation for running the pipeline, generating manuscript figures, and building/evaluating machine learning classifiers are available on GitHub [https://github.com/SpencerEricksen/PubChem_DataMining_OXPHOS].

Declarations

Ethics approval and consent to participate

Not applicable as human subjects and vertebrate animals were not included in this study.

Competing interests

The authors declare no competing interests.

Author details

[1]University of Wisconsin-Madison, Department of Obstetrics and Gynecology, Madison, WI 53705, USA. [2]Department of Obstetrics and Gynecology, Qilu Hospital of Shandong University, Jinan, Shandong 250012, People's Republic of China. [3]University of Wisconsin-Madison, UW-Carbone Cancer Center, Drug Development Core, Small Molecule Screening Facility, Wisconsin Institutes for Medical Research, 1111 Highland Avenue, Madison, WI 53705, USA.

Received: 25 April 2024 Accepted: 15 September 2024
Published online: 07 October 2024

References

1. Pinto GP, Hendrikse NM, Stourac J, Damborsky J, Bednar D (2022) Virtual screening of potential anticancer drugs based on microbial products. Semin Cancer Biol 86:1207–1217
2. da Silva Rocha SFL, Olanda CG, Fokoue HH, Sant'Anna CMR (2019) Virtual screening techniques in drug discovery: review and recent applications. Curr Top Med Chem 19:1751–1767
3. Shaker B, Ahmad S, Lee J, Jung C, Na D (2021) In silico methods and tools for drug discovery. Comput Biol Med 137:104851
4. Ballante F, Kooistra AJ, Kampen S, de Graaf C, Carlsson J (2021) Structure-based virtual screening for ligands of G protein-coupled receptors: what can molecular docking do for you? Pharmacol Rev 73:527–565
5. Dalkıran A, Atakan A, Rifaioğlu AS et al (2023) Transfer learning for drug-target interaction prediction. Bioinformatics 39:i103–i110
6. Cai C, Wang S, Xu Y et al (2020) Transfer learning for drug discovery. J Med Chem 63:8683–8694
7. Salem M, Khormali A, Arshadi AK, Webb J, Yuan J-S (2020) TranScreen: transfer learning on graph-based anti-cancer virtual screening model. Big Data Cogn Comput 4:16
8. Kim S, Chen J, Cheng T et al (2022) PubChem 2023 update. Nucleic Acids Res 51:D1373–D1380
9. Mendez D, Gaulton A, Bento AP et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930-d940
10. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2015) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053
11. Kapur A, Mehta P, Simmons AD et al (2022) Atovaquone: an inhibitor of oxidative phosphorylation as studied in gynecologic cancers. Cancers (Basel) 14:2297
12. Nayak AP, Kapur A, Barroilhet L, Patankar MS (2018) Oxidative phosphorylation: a target for novel therapeutic strategies against ovarian cancer. Cancers (Basel) 10:337
13. Xu Y, Xue D, Bankhead A III, Neamati N (2020) Why all the fuss about oxidative phosphorylation (OXPHOS)? J Med Chem 63:14276–14307
14. García Rubiño ME, Carrillo E, Ruiz Alcalá G, Domínguez-Martín A, Marchal JA, Boulaiz H (2019) Phenformin as an anticancer agent: challenges and prospects. Int J Mol Sci 20:3316
15. Yap TA, Daver N, Mahendra M et al (2023) Complex I inhibitor of oxidative phosphorylation in advanced solid tumors and acute myeloid leukemia: phase I trials. Nat Med 29:115–126
16. Machado ND, Heather LC, Harris AL, Higgins GS (2023) Targeting mitochondrial oxidative phosphorylation: lessons, advantages, and opportunities. Br J Cancer 129:897–899
17. Kapur A, Ayuso JM, Rehman S et al (2023) Oxidative phosphorylation inhibitors inhibit proliferation of endometriosis cells. Reproduction 165:617–628
18. Kapur A, Beres T, Rathi K et al (2018) Oxidative stress via inhibition of the mitochondrial electron transport and Nrf-2-mediated anti-oxidative response regulate the cytotoxic activity of plumbagin. Sci Rep 8:1073
19. Kapur A, Felder M, Fass L et al (2016) Modulation of oxidative stress and subsequent induction of apoptosis and endoplasmic reticulum stress allows citral to decrease cancer cell proliferation. Sci Rep 6:27530
20. Zeng S, Kapur A, Patankar MS, Xiong MP (2015) Formulation, characterization, and antitumor properties of trans- and cis-citral in the 4T1 breast cancer xenograft mouse model. Pharm Res 32:2548–2558

21. Liu Y, Whelan RJ, Pattnaik BR et al (2012) Terpenoids from *Zingiber officinale* (Ginger) induce apoptosis in endometrial cancer cells through the activation of p53. PLoS ONE 7:e53178
22. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37:W623-633
23. Alharbi Y, Kapur A, Felder M et al (2019) Plumbagin-induced oxidative stress leads to inhibition of Na(+)/K(+)-ATPase (NKA) in canine cancer cells. Sci Rep 9:11471
24. Breiman L (2001) Random forests. Mach Learn 45:5–32
25. Permutation Importance vs Random Forest Feature Importance (MDI). https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html.
26. Chance B, Hollunger G (1963) Inhibition of electron and energy transfer in mitochondria: I. Effects of amytal, thiopental, rotenone, progesterone, and methylene glycol. J Biol Chem 238:418–431
27. Naguib A, Mathew G, Reczek CR et al (2018) Mitochondrial complex I inhibitors expose a vulnerability for selective killing of Pten-null cells. Cell Rep 23:58–67
28. Brunmair B, Lest A, Staniek K et al (2004) Fenofibrate impairs rat mitochondrial function by inhibition of respiratory complex I. J Pharmacol Exp Ther 311:109–114
29. Miyadera H, Shiomi K, Ui H et al (2003) Atpenins, potent and specific inhibitors of mitochondrial complex II (succinate-ubiquinone oxidoreductase). Proc Natl Acad Sci 100:473–477
30. Hemmerich J, Troger F, Füzi B, Ecker GF (2020) Using machine learning methods and structural alerts for prediction of mitochondrial toxicity. Mol Inform 39:e2000005
31. Riss TL, Moravec RA, Niles AL (2011) Cytotoxicity testing: measuring viable cells, dead cells, and detecting mechanism of cell death. Methods Mol Biol 740:103–114
32. Ertl P, Roggo S, Schuffenhauer A (2008) Natural product-likeness score and its application for prioritization of compound libraries. J Chem Inf Model 48:68–74
33. Erlanson DA (2015) Learning from PAINful lessons. J Med Chem 58:2088–2090
34. Baell JB, Nissink JWM (2018) Seven Year itch: pan-assay interference compounds (PAINS) in 2017-utility and limitations. ACS Chem Biol 13:36–44
35. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754
36. Huggins JE, Wren PA, Gruis KL (2011) What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. Amyotroph Lateral Scler 12:318–324
37. Alnammi M, Liu S, Ericksen SS et al (2023) Evaluating scalable supervised learning for synthesize-on-demand chemical libraries. J Chem Inf Model 63:5513–5528
38. McInnes L, Healy J. UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv. 2018; http://arxiv.org/abs/1802.03426.

Publisher's Note