scientific data

OPEN

DATA DESCRIPTOR

An improved dataset of force fields, electronic and physicochemical descriptors of metabolic substrates

Alessio Macorano[1], Angelica Mazzolari[1], Giuliano Malloci[2], Alessandro Pedretti[1], Giulio Vistoli[1] & Silvia Gervasoni[2] ✉

In silico **prediction of xenobiotic metabolism is an important strategy to accelerate the drug discovery process, as candidate compounds often fail in clinical phases due to their poor pharmacokinetic profiles. Here we present Meta^{QM}, a dataset of quantum-mechanical (QM) optimized metabolic substrates, including force field parameters, electronic and physicochemical properties. Meta^{QM} comprises 2054 metabolic substrates extracted from the MetaQSAR database. We provide QM-optimized geometries, General Amber Force Field (FF) parameters for all studied molecules, and an extended set of structural and physicochemical descriptors as calculated by DFT and PM7 methods. The generated data can be used in different types of analysis. FF parameters can be applied to perform classical molecular mechanics calculations as exemplified by the validating molecular dynamics simulations reported here. The calculated descriptors can represent input features for developing improved predictive models for metabolism and drug design, as exemplified in this work. Finally, the QM-optimized molecular structures are valuable starting points for both ligand- and structure-based analyses such as pharmacophore mapping and docking simulations.**

Background & Summary

The prediction of drug metabolism has been attracting great interest in recent years for its capacity to rapidly screen huge databases of compounds allowing a cost-effective discarding of the molecules with a predicted unfavourable profile. Notably, such an *in silico* screening can be performed in the early phases of the drug discovery process with clear benefits in the reduction of the failures related to pharmacokinetic and toxicological concerns[1].

The approaches for metabolism prediction can be subdivided into two major groups. On one hand, the local methods focus on a specific metabolic reaction and on the related metabolizing enzyme(s). On the other hand, global methods aim to predict the overall metabolic fate a given compound can undergo. Even though the global approaches often involve knowledge-based metabolic rules, local and global methods can develop their predictive models by exploiting both ligand- and structure-based approaches[2]. Over the last years, all metabolism predictive studies greatly benefit from the artificial intelligence algorithms which allow the predictive performances to be constantly enhanced[3].

The major factor which has so far limited the development of metabolism predictive models (especially involving global methods) is the scarcity of highly accurate and extended datasets. Most available metabolic resources are indeed collected by automatic interrogation of other databases[4] combining xenobiotic and endogenous metabolic data for omics analyses[5]. Hence, we recently proposed the MetaQSAR resource[6], a manually curated database collected by meta-analysis of the recent primary specialized literature. MetaQSAR comprises 3788 first generation metabolic reactions which are grouped by a finely organized classification which subdivides them in 3 major classes, 21 classes and 101 subclasses[7]. MetaQSAR is thus a fruitful source of highly accurate datasets well suited for developing metabolism predictive analyses which indeed proved successful in both local[3] and global ligand-based studies[8,9]. Altogether, the developed predictive models emphasized the key role of electronic descriptors, a quite expected outcome when considering their capacity to parameterize the intrinsic reactivity of each atom/molecule. The hitherto published studies involved electronic descriptors as computed by

[1]Dipartimento di Scienze Farmaceutiche, Università degli Studi di Milano, via Mangiagalli 25, 20133, Milano, Italy. [2]Dipartimento di Fisica, Università degli Studi di Cagliari, Cittadella Universitaria, S.P. Monserrato-Sestu Km 0.7, I-09042, Monserrato, CA, Italy. ✉e-mail: silvia.gervasoni@dsf.unica.it

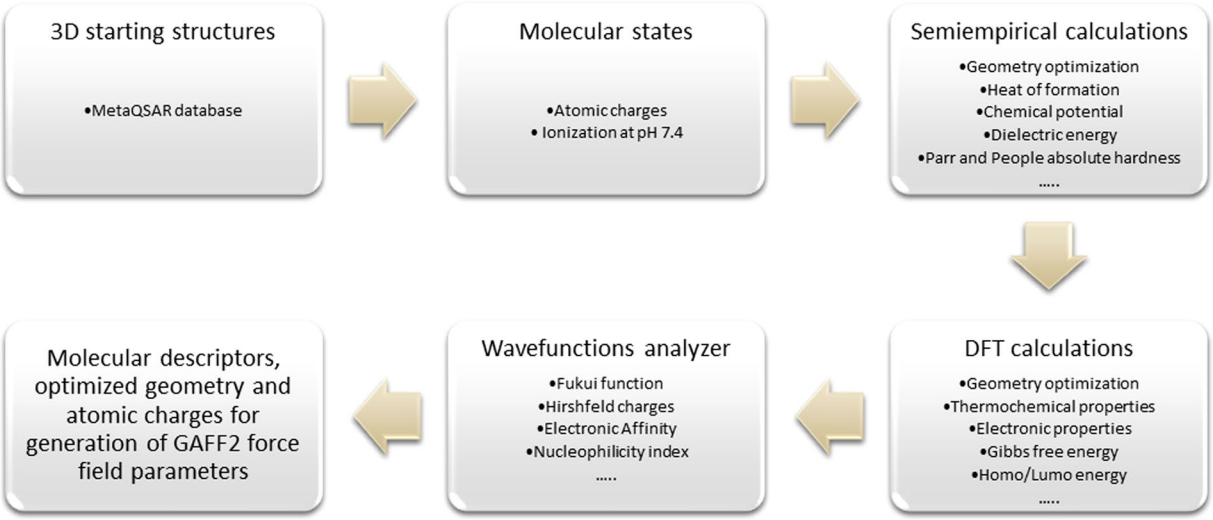

Fig. 1 Schematic view of the computational workflow adopted in this work.

semi-empirical methods, an almost compulsory choice to reduce the computational costs[10]. Nevertheless, one may imagine that the predictive power of these descriptors should parallel the level of theory by which they are calculated.

Hence, we undertook a highly demanding campaign of DFT calculations in which all the 2054 first generation substrates, as extracted from MetaQSAR, underwent DFT-based full optimization and frequency calculations. Here, we release all the so derived molecular data for more than 2000 molecules including: (a) two datasets of all the QM-optimized substrates (at both DFT and semiempirical PM7 levels, with the corresponding Gaussian output files); (b) an homogeneous database of General Amber Force-Field parameters including several compounds bearing non-standard atoms; (c) all the derived electronic descriptors; (d) an extended set of physicochemical descriptors as computed by using the DFT-optimized conformations.

By considering the structural richness of the simulated molecules, the present data can have many applications. First, the collected force-field parameters can be used to perform molecular mechanics calculations as exemplified by the here reported validating molecular dynamics runs on the compounds including non-standard atoms. Second, as exemplified by few selected test-cases, the computed descriptors can be utilized to develop improved predictive models (not necessarily focused on drug metabolism). Third, the QM-optimized structures can represent valuable starting points for various ligand- and structure-based studies. Notice that the collected dataset of optimized structures mostly comprises marketed drugs and drug-like molecules and, therefore, it can be particularly suited for repurposing and virtual screening campaigns.

Methods

DFT-optimization of MetaQSAR molecules. As schematized in the workflow of Fig. 1, the 3D structures of the first-generation substrates contained in the MetaQSAR database[7] (overall 2054 molecules) were generated at physiological pH 7.4 by the VEGA[11] program.

All compounds underwent a two steps geometry optimization, using first a semi-empirical and then a Density Functional Theory (DFT)[12] level of theory. In detail, the semi-empirical calculations were performed using the MOPAC 2016 software[13] and the PM7[14] Hamiltonian. The DFT calculations included full optimization and frequency calculation using the Gaussian 16 software (Revision A.03)[15]. The hybrid B3LYP functional[16] is widely recognized as the standard for the systematic study of organic molecules[17,18]. It has been used in combination with the 6–31 G* basis set for C and H, and 6–31 + G* for heteroatoms such as N, O, P, S. For compounds containing "non-standard" atoms (*i.e.*, Pt, As, Hg, Se, Pb), LANL2DZ[19] effective core potential (ECP) and double zeta basis set were used. In all cases, the absence of imaginary frequency modes for the optimized structure of the ligand confirms a true minimum on the potential energy surface. At the optimized geometry, the Multiwfn 3.8 program[20] was used to calculate Hirshfeld Charges[21] within the conceptual density functional theory (CDFT)[22]. Hirshfeld population analysis (HPA) has proven to be a suitable choice compared to other population analysis schemes[23,24]. It is particularly effective for studying and obtaining Fukui functions[25], dual descriptors and Hirshfeld charges[21] itself, which reveals nucleophile and/or electrophile reactive centers of the ligand that underwent a metabolic reaction. GuassSum[26] 3.0 was used to extract all the information associated with each molecular orbital, from the previously generated output files.

To check the quality of DFT calculations, we compared the DFT optimized structures with experimental crystallographic structures retrieved from the Cambridge Structural Database (CSD)[27] of the Cambridge Crystallographic Data Center (CCDC). In detail, a subset of 100 molecules were selected from the MetaQSAR database considering their structural diversity, by means of RDkit diversity picker[28] as implemented in the KNIME 4.6.4 analytic platform[29]. We restricted the selection to CSD experimental structures with R% factor value < 5. The root mean square deviation (RMSD) values on heavy atoms, between DFT optimized structures

and experimental structures were calculated by using the Visual Molecular Dynamics software (VMD)[30] along with the corresponding $RMSD_w$ (average value) (see Technical Validation section).

Amber force-field parameters generation. General Amber Force-Field parameters (GAFF2)[31] were generated starting from the Gaussian log files and assigning the Hirshfeld atomic charges obtained as described above. For compounds containing non-standard atoms (*i.e.*, selenium. platinum, arsenic, iron, silicon, mercury, tin, and boron) we generated bonded parameters following the metal center parameter builder procedure (MCPB.py)[32] as implemented in Amber22[33]. For compounds containing boron atoms, not supported by the MCPB.py procedure, we used the parameters reported by Tafi *et al.*[34]. For molecules containing non-standard atoms the quality of the GAFF2 parameters was checked by performing a molecular mechanics optimization using the conjugated gradient algorithm followed by a 100 ns-long molecular dynamics simulation in explicit water solution using Amber22. In detail, compounds were inserted into a box of OPC water molecules[35] and the systems were neutralized by adding either Na^+ or Cl^- counter ions. The hydrogen mass repartition scheme was adopted[36], as well as the SHAKE algorithm[37]. The NPT production runs were preceded by an energy minimization, a heating followed by a cooling phase, as described previously[38]. We used a time step of 4 fs, a cutoff for non-bonded interaction of 9 Å, the Langevin thermostat and the Berendsen barostat for keeping the temperature at 310 K and the pressure at 1 Atm. Periodic boundary conditions and PME method were applied.

Metabolism prediction model building. To show how the molecular descriptors provided by our study can be helpful in predicting the metabolism of compounds, we built machine learning models to predict whether a compound undergoes three selected metabolic reactions: glutathione or generic sulfur conjugation (MetaQSAR class 24), hydrolysis of amides, lactams, and peptides (MetaQSAR class 12), and oxidation and reduction of sulfur atoms (MetaQSAR class 08). A binary classification model based on the MetaQSAR system was used. The program Weka 3.8.6[39] was used to build the model, using the Random Forest algorithm with the following parameters: (1) batch size = 100; (2) number of threads = 1; (3) number of iterations = 100; (4) the attribute importance was not evaluated. The most significant features were selected by using the Weka program according to both the BestFirst search algorithm (direction = Forward; lookupCacheSize = 1; searchTermination = 5) and theWrapperSubsetEval attribute evaluator (classifier = RandomForest with default settings; doNotCheckCapabilities = False; evaluationMeasure = accuracy,RMSE; folds = 5; seed = 1; threshold = 0.01). The performance of the models was evaluated using different metrics: Precision and Recall, see Eqs. (1–5)), Matthew's Correlation Coefficient (MCC), and the Receiver Operating Characteristic Curve Area (ROC Area)[40], obtained through a 10-fold cross-validation. Specifically, in the following equations, each symbol is represented as follows: TP for true positive, TN for true negative, FP for false positive, and FN for false negative.

$$\text{MCC: } \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

$$\text{Precision Class\_YES: } \frac{TP}{TP + FP} \quad (2)$$

$$\text{Precision Class\_NO: } \frac{TN}{TN + FN} \quad (3)$$

$$\text{Recall Class\_YES: } \frac{TP}{TP + FN} \quad (4)$$

$$\text{Recall Class\_NO: } \frac{TN}{TN + FP} \quad (5)$$

Data Records

The $Meta^{QM}$ database is available on figshare[41]. Figure 2 shows the structure of the database.

We shared two comma separated files containing molecular descriptors as derived from semiempirical optimized structures (*MetaQM_PM7-based_descriptors.csv*) and from DFT optimized structures (*MetaQM_DFT-based_descriptors.csv*). The list of computed descriptors together with a precise description of their meaning is reported in Supporting Information (Table S1). The DFT and PM7 optimized structures are contained in two MOL2 database files (*MetaQM-DFT_opt_structures.mol2* and *MetaQM-PM7_opt_structures.mol2*).

The compress file *MetaQM.tar.gz* contains 2054 folders, one for each compound; the folder and the included files are named after the compound (*e.g.*, Abacavir/). Each directory includes 6 files: the Gaussian output of DFT calculations (*e.g.*, *Abacavir.log*), the list of atomic charges computed at DFT-level of theory (*e.g.*, *Abacavir.txt*), the list of all molecular orbitals (*e.g.*, *Abacavir.mo*), the .mol2 file used for the force field generation (*e.g.*, *Abacavir.mol2*), and the two GAFF2 files (*e.g.*, *Abacavir.prepin*, *Abacavir.frcmod*). For compounds containing non-standard atoms (Table S2) the .lib file is supplied instead of .prepin (*e.g.*, *Arsenate.lib*). For the Ferroquine compound, only the .log and .txt files are supplied (see Methods).

For compounds with non-standard atoms an additional subfolder (MD_traj/) is further provided. MD_traj/ contains the topology and coordinates (*e.g.*, *Arsenate.prmtop*, *Arsenate.inpcrd*) of the solvated compound used

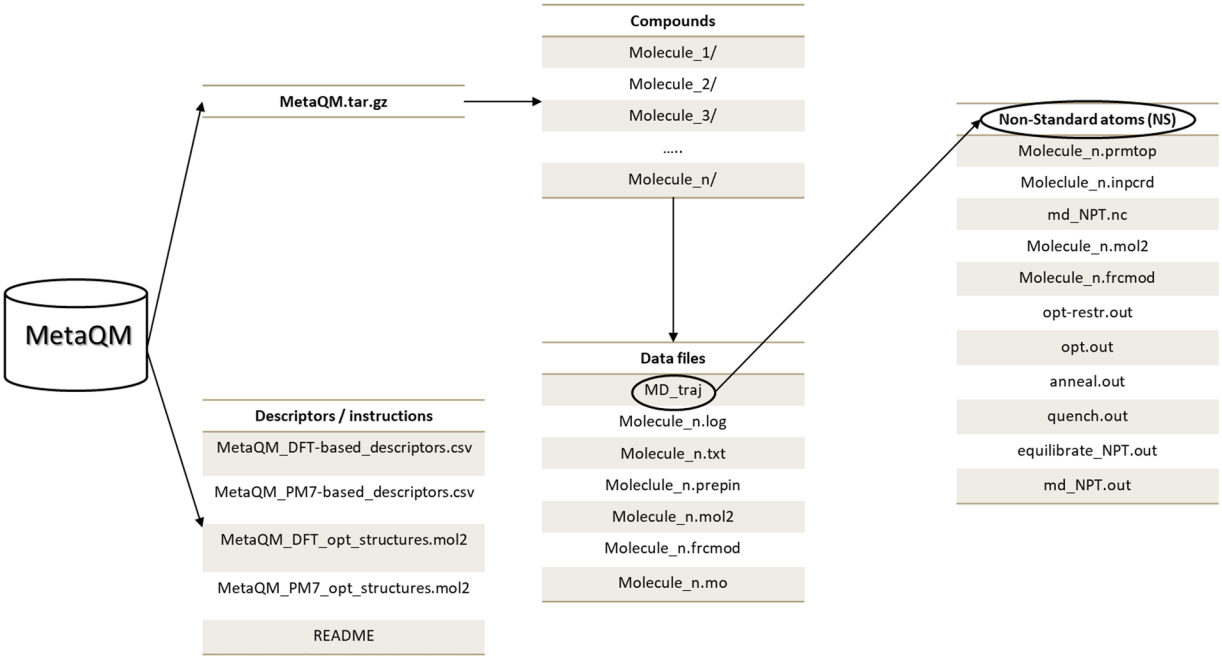

Fig. 2 Schematic representation of Meta^{QM} structure.

Compound	RMSD (Å)	Formula	Molecular weight	Atoms
Ar-67	0.76	$C_{26}H_{30}N_2O_5Si$	478.61	64
Arsenate	0.44	HO_4As	139.93	6
ARSENITE	0.61	H_3O_3As	125.94	7
Bortezomib	0.83	$C_{19}H_{25}N_4O_4B$	384.24	53
Carboplatin	0.47	$H_6N_2Cl_2Pt$	300.04	11
Cisplatin	0.41	$C_2H_7O_2As$	137.99	12
Dimethyl arsinate	0.76	C_2H_6Sn	148.78	9
Dimethyltin	0.40	$C_{11}H_{17}NO_4B$	238.07	34
GSK2251052	1.13	$C_{12}H_{15}N_6OS_2As$	398.34	37
Melarsoprol	0.76	Cl_2Hg	271.50	3
Mercury chloride	0.06	CH_3Hg	215.62	5
Methylmercury	0.01	CH_3ClHg	251.08	6
Methylmercury chloride	0.03	$C_8H_{14}N_2O_4Pt$	397.29	29
Oxaliplatin	0.05	$C_5H_{11}NO_2Se$	196.11	20
Seleno-L-methionine	0.50	$C_{20}H_{24}NO_2FClSi$	392.95	50
Sila-Haloperidol	0.56	$C_6H_{12}N_2O_4Pt$	371.25	25
Average	**0.49**			

Table 1. RMSD values (Å) as computed on all atoms between molecular mechanics and DFT-optimized structure of compounds containing non-standard atoms.

for the validating MD simulation. The MD trajectory is contained in the md_*NPT.nc* file. The Amber22 output files from minimization, equilibration and production steps are supplied as well *(opt-restr.out*, *opt.out*, *anneal.out*, *quench.out*, *equilibrate_NPT.out*, *md_NPT.out*).

Technical Validation

DFT-optimized structures. The validation of molecular structures optimized by DFT calculations was carried out by selecting 100 structurally diverse molecules from the simulated MetaQSAR substrates. We restricted the selection to the experimental crystal structures deposited on CSD with high quality resolution (*i.e.*, R% < 5). In the case of compounds with more available structures, we chose the one with the lowest R% value.

The resolved structures of the so selected 100 compounds were then compared with the corresponding DFT optimized conformations. For each selected compound, Table S3 compiles the reference CSD code and R-factor together with the resulting RMSD value. The RMSD mean value ($RMSD_w$) is also reported.

Almost 70% of the molecules have RMSD values < 1 Å, indicating that the DFT optimized structures are in good agreement with the corresponding resolved structures, with the lowest value being 0.01 Å for

	n	MCC	ROC Area	Precision YES	Precision NO	Recall YES	Recall NO
Class 24	338	0.71	0.92	0.87	0.84	0.83	0.88
Class 12	234	0.41	0.75	0.73	0.68	0.64	0.77
Class 08	254	0.68	0.89	0.85	0.83	0.82	0.86

Table 2. Performances of the three machine-learning metabolic predictions based on Meta^{QM} descriptors. The MCC value ranges from −1 (worse) to 1 (best), all the other metrics range from 0 (worse) to 1 (best). The overall number of instances for each model is also reported (n).

Class 24	Class 12	Class 08
ChiralAtms	EzBnds	Mass
HbDon	ChiralAtms	HbDon
Impropers	Psa	ChiralAtms
Rings	Impropers	Psa
Hirshfeld_positive_charges	VirtualLogP	**Gap**
piS_TOTAL	**Hirshfeld_positive_charges**	**Chemical_potential**
	Fukui_positive	**Ionization_potential**
	D.E_Total_PM7	**Nucleophilicity_index**
		Thermal_energy
		Fukui_negative

Table 3. Features selection for the three test-case metabolic predictive models (see Table S1 for details about each molecular descriptor). QM-based electronic properties are highlighted in bold.

Tetrafluoroethene, Coumarin and Dioxane. In contrast, the 9% of the cases show large structural difference with RMSD > 2 Å, the maximum value of 3.49 Å being observed for Dabrafenib. As expected, the obtained results suggest that flexible molecules give rise to high RMSD values, while rigid molecules reveal low RMSD values. However, the RMSD mean value of 0.76 Å confirms an overall agreement between the DFT optimized conformations and the experimental structures.

The molecular descriptors were computed on the DFT optimized conformations, that can differ in the general case from the conformations of compounds when in complex with metabolic enzymes. Therefore, to check the robustness of the provided dataset, we collected a subset of 20 diverse representative compounds (ranging from 12 to 50 heavy atoms, and 0 to 18 flexible torsions), for which the experimental structures in complex with metabolic enzymes are available in the Protein Data Bank[42]. We then computed the QM-based descriptors (with both DFT and PM7 methods) on the experimental conformation, and we compared the results with those derived from the corresponding QM-optimized geometries (Supporting_TableS 4a, Supporting_TableS 4b). We obtained small differences between the different series of descriptors, with average percentage variations of 11% for the PM7-based descriptors, and 6% for DFT-based descriptors, indicating the overall reliability of the data.

Amber force-field parameters. To validate the quality of the GAFF2 parameterization for molecules containing non-standard atoms, we compared the optimized geometries derived from the molecular mechanics minimization with those obtained by DFT calculations. Table 1 shows the RMSD values between the two structures, computed on all atoms. In 9 cases out of 16, the two compared structures are almost identical (*i.e.*, RMSD < 0.5 Å) and only one molecule shows a RMSD value greater than 1 Å. Overall, the average value considering all "non-standard" cases is equal to 0.49 Å thus confirming the reliability of the computed force field parameters. For the same molecules, the force field parameters were utilized to perform 100 ns-long MD simulations in explicit water solution to further test the reliability of the bonded parameters. The visual inspection of the MD trajectories, available on figshare, reveals a satisfactory stability of distances/angles/torsions involving non-standard atoms along the 100 ns timescale, thus demonstrating the reliability of the corresponding bonded parameters.

Example of metabolic predictions using Meta^{QM}. To test how the Meta^{QM} descriptors can feed predictive machine learning models of metabolism, we performed three tests of selected metabolic predictions. Specifically, we followed the MetaQSAR metabolic reaction classification system to predict whether compounds undergo: (1) glutathione conjugation (metabolic class 24, 169 substrates plus 169 non-substrates), (2) hydrolysis (metabolic class 12, 117 substrates plus 117 non-substrates), and (3) oxidation and reduction of sulphur atoms (metabolic class 08, 127 substrates plus 127 non-substrates). For each prediction model, we used a balanced data set consisting of 50% of molecules that undergo the reaction (substrates) and 50% of molecules that do not undergo the reaction (non-substrates). To highlight the role of electronic descriptors, each reactive functional group was also used for non-substrate species. The results of the prediction models are shown in Table 2.

Overall, the performance of the three prediction models is satisfactory, with the glutathione conjugation reaction (class 24) and oxidations of sulphur atoms (class 08) showing the best results with an MCC of 0.71 and 0.68, respectively. The ROC curves (*i.e.*, true positive rate (TPR) vs False Positive Rate (FPR)) for all the three classes are reported in Figure S1. Class 12, representing the hydrolysis of amides, lactams and peptides, obtained

a lower but acceptable performance in terms of prediction. These results could be related to the MetaQSAR classification scheme of reactions, for which both conjugation reactions and oxidation on sulphur atoms (classes 24 and 08) include more homogeneous metabolic reactions. Instead, for class 12 (hydrolysis of amides, lactams and peptides), the collected metabolic reactions are more heterogeneous, which may partly explain the lower but acceptable performance of the corresponding model. Although the samples used to build each model included the same reactive functional group for both substrates and non-substrates, that can make the prediction more difficult, the novel electronic descriptors presented here show an overall satisfactory performance.

All predictive models contain three types of molecular descriptors (*i.e.*, phys-chem, DFT-based, and semiempirical) (Table 3) obtained after 10-fold cross validation (Table 3).

Class 08 is characterized by the highest number of features with respect to the other two classes, especially considering the electronic parameters. This could be ascribed to the complex biochemical mechanisms of the oxidation reaction catalyzed by CYP450 on the sulfur atom. In detail, the Cdp I complex is also referred to as an "electrophilic oxidant"[43], which could explain why both *Fukui_negative* and *Nucleophilicity_Index*, both capturing atomic and molecular nucleophilicity, are identified as important features for this model. In addition, other electronic descriptors are identified, that globally describe the chemical reactivity of the oxidation reaction. The physicochemical parameters are related to molecular size and shape except for *HbDon* and *Psa*, which encode both polarity and the presence of chemical groups susceptible to metabolism. When considering the other two classes, the number of electronic features is lower, possibly due to the simpler reaction mechanisms compared to the previous one. In these cases, electronic parameters that encode both electrophilicity and chemical reactivity (*Hirshfeld_positive_charges*, *Fukui_positive*, *D.E_Total_PM7*, and *piS_TOTAL*) are found to be important, as well as physicochemical parameters accounting for molecular size, molecular shape and polarity/lipophilicity properties.

Code availability

The starting 3D structures of compounds were retrieved from the MetaQSAR[7] database (available under licence). The ionization of compounds was performed using VEGA 3.2.1[11]. The software MOPAC[13] 2016 was used for the semiempirical optimizations and for the calculation of semiempirical descriptors. Gaussian16[15] (Revision A.03) was employed for both the DFT-based geometry optimizations and the descriptors collection, and Multiwfn 3.8[20] was used only for the DFT-based descriptor computation. For each compound, GaussSum 3.0[26] was used to extract the molecular orbitals from the Gaussian16 output files in combination with a personalized script (*extract_orbitals.py*). VMD 1.9.4[30] was used for the visualization and computation of RMSD of compounds. Amber22[33] was used for the generation of the force field parameters and the MD simulations. Weka 3.8.6[39] was employed to create the metabolism predictive machine learning models.

Received: 20 March 2024; Accepted: 30 July 2024;
Published online: 27 August 2024

References

1. Kazmi, S. R., Jun, R., Yu, M. S., Jung, C. & Na, D. *In silico* approaches and tools for the prediction of drug metabolism and fate: A review. *Comput. Biol. Med.* **106**, 54–64 (2019).
2. Kirchmair, J. *et al*. Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Discov.* **14**, 387–404 (2015).
3. Dudas, B. & Miteva, M. A. Computational and artificial intelligence-based approaches for drug metabolism and transport prediction. *Trends Pharmacol. Sci.* **45**, 39–55 (2024).
4. Karp, P. D. Can we replace curation with information extraction software? *Database* **2016**, baw150 (2016).
5. Wishart, D. S. *et al*. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).
6. Testa, B., Pedretti, A. & Vistoli, G. Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Discov. Today* **17**, 549–560 (2012).
7. Pedretti, A., Mazzolari, A., Vistoli, G. & Testa, B. MetaQSAR: An Integrated Database Engine to Manage and Analyze Metabolic Data. *J. Med. Chem.* **61**, 1019–1030 (2018).
8. Chen, Y. *et al*. Active Learning Approach for Guiding Site-of-Metabolism Measurement and Annotation. *J. Chem. Inf. Model.* **64**, 348–358 (2024).
9. Mazzolari, A. *et al*. MetaSpot: A General Approach for Recognizing the Reactive Atoms Undergoing Metabolic Reactions Based on the MetaQSAR Database. *Int. J. Mol. Sci.* **24**, 11064 (2023).
10. Mazzolari, A., Scaccabarozzi, A., Vistoli, G. & Pedretti, A. MetaClass, a Comprehensive Classification System for Predicting the Occurrence of Metabolic Reactions Based on the MetaQSAR Database. *Molecules* **26**, 5857 (2021).
11. Pedretti, A., Mazzolari, A., Gervasoni, S., Fumagalli, L. & Vistoli, G. The VEGA suite of programs: an versatile platform for cheminformatics and drug design projects. *Bioinformatics* **37**, 1174–1175 (2021).
12. Kohn, W. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Rev. Mod. Phys.* **71**, 1253 (1999).
13. Stewart, J. J. P. MOPAC2016. (216AD).
14. Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **19**, 1–32 (2013).
15. Frisch, M. J. *et al*. Gaussian 16, Gaussian, Inc., Wallingford CT. Revision A.03 (2016).
16. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).
17. Tirado-Rives, J. & Jorgensen, W. L. Performance of B3LYP density functional methods for a large set of organic molecules. *J. Chem. Theory Comput.* **4**, 297–306 (2008).
18. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. General Performance of Density Functionals. *J. Phys. Chem. A* **111**, 10439–10452 (2007).
19. Hay, P. J. & Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.* **82**, 299–310 (1985).
20. Lu, T. & Chen, F. Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **33**, 580–592 (2012).
21. Liu, S., Rong, C. & Lu, T. Information conservation principle determines electrophilicity, nucleophilicity, and regioselectivity. *J. Phys. Chem. A* **118**, 3698–3704 (2014).
22. Domingo, L. R., Ríos-Gutiérrez, M. & Pérez, P. Applications of the Conceptual Density Functional Theory Indices to Organic Chemistry Reactivity. *Molecules* **21**, 748 (2016).

23. Roy, R. K. Stockholders Charge Partitioning Technique. A Reliable Electron Population Analysis Scheme to Predict Intramolecular Reactivity Sequence. *J. Phys. Chem. A* **107**, 10428–10434 (2003).
24. Wang, B., Rong, C., Chattaraj, P. K. & Liu, S. A comparative study to predict regioselectivity, electrophilicity and nucleophilicity with Fukui function and Hirshfeld charge. *Theor. Chem. Acc.* **138**, 1–9 (2019).
25. Oláh, J. & Alsenoy, C. Van & Sannigrahi, A. B. Condensed Fukui Functions Derived from Stockholder Charges: Assessment of Their Performance as Local Reactivity Descriptors. *J. Phys. Chem. A* **106**, 3885–3890 (2002).
26. O'Boyle, N. M., Tenderholt, A. L. & Langner, K. M. Cclib: A library for package-independent computational chemistry algorithms. *J. Comput. Chem.* **29**, 839–845 (2008).
27. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016).
28. RDKit: Open-source cheminformatics. https://www.rdkit.org.
29. Berthold, M. R. *et al.* KNIME: The Konstanz information miner. *Stud. Classif. Data Aal. Knowl. Organ*. 319–326, https://doi.org/10.1007/978-3-540-78246-9_38/COVER (2008).
30. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
31. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
32. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.* **56**, 599–604 (2016).
33. Case, D. A. *et al.* University of California, San Francisco (2022).
34. Tafi, A. *et al.* AMBER force field implementation of the boronate function to simulate the inhibition of β-lactamases by alkyl and aryl boronic acids. *Eur. J. Med. Chem.* **40**, 1134–1142 (2005).
35. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building water models: A different approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).
36. Jung, J. *et al.* Optimized Hydrogen Mass Repartitioning Scheme Combined with Accurate Temperature/Pressure Evaluations for Thermodynamic and Kinetic Properties of Biological Systems. *J. Chem. Theory Comput.* **17**, 5312–5321 (2021).
37. Kräutler, V., Gunsteren, W. Fvan & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22**, 501–508 (2001).
38. Gervasoni, S. *et al.* AB-DB: Force-Field parameters, MD trajectories, QM-based data, and Descriptors of Antimicrobials. *Sci. Data 2022 91* **9**, 1–12 (2022).
39. Witten, I. H., Frank, E. & Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. *Data Mining: Practical Machine Learning Tools and Techniques* https://doi.org/10.1016/C2009-0-19715-5 (Elsevier, 2011).
40. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
41. Macorano, A. *et al.* An improved dataset of force fields, electronic and physicochemical descriptors of metabolic substrates, *figshare*, https://doi.org/10.6084/m9.figshare.24574495 (2024).
42. Burley, S. K. *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
43. Shaik, S. *et al.* The 'Rebound Controversy': An Overview and Theoretical Modeling of the Rebound Step in C-H Hydroxylation by Cytochrome P450. *Eur. J. Inorg. Chem*. 207–226, https://doi.org/10.1002/EJIC.200300448 (2004).

Acknowledgements

S.G. and G.M. gratefully acknowledge the Health Extended ALliance for Innovative Therapies, Advanced Lab-research, and Integrated Approaches of Precision Medicine partnership (HEAL ITALIA), founded by the Italian Ministry of University and Research, PNRR, mission 4, component 2, investment 1.3, project number PE00000019 (University of Cagliari). All the authors gratefully acknowledge the support from the University of Milan through the institutional APC initiative.

Author contributions

G.V., A.P. and G.M. conceived the project, A.Mac. and S.G. performed the calculations, A.Maz. collected the dataset, all authors wrote and reviewed the manuscript.

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/s41597-024-03707-0.

Correspondence and requests for materials should be addressed to S.G.