High-throughput screening and prediction models building for novel hemozoin inhibitors using physicochemical properties

Nguyen Tien Huy,[1,2,3*] Pham Lan Chi,[2,4] Jun Nagai,[2,5] Tran Ngoc Dang,[6,7] Evaristus Chibunna Mbanefo,[2,3,4] Ali Mahmoud Ahmed,[8] Nguyen Phuoc Long,[3,6] Le Thi Bich Thoa,[3,6] Le Phi Hung,[3,6] Titouna Afaf,[2,4] Kaeko Kamei,[9] Hiroshi Ueda,[2,5] Kenji Hirayama[2,4*]

Department of Clinical Product Development, Institute of Tropical Medicine (NEKKEN), Nagasaki University, Sakamoto, Nagasaki, Japan[1]; Graduate School of Biomedical Sciences, Nagasaki University, Bunkyo-machi, Nagasaki, Japan[2]; Online Research Club, Nagasaki University, Sakamoto, Nagasaki, Japan[3]; Department of Immunogenetics, Institute of Tropical Medicine (NEKKEN), Nagasaki University, Sakamoto, Nagasaki, Japan[4]; Department of Pharmacology and Therapeutic Innovation, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki University, Nagasaki, Japan[5]; University of Medicine and Pharmacy at Ho Chi Minh City, Hong Bang, District 5, Ho Chi Minh, Vietnam[6]; Department of Health Care Policy and Management, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, Japan[7]; Faculty of Medicine, Al-Azhar University, Cairo, Egypt[8]; Department of Biomolecular Engineering, Kyoto Institute of Technology, Sakyo-ku, Japan[9].

Running title: Prediction models of novel hemozoin inhibitors

* Address correspondence to Nguyen Tien Huy, tienhuy@nagasaki-u.ac.jp; or Kenji Hirayama, hiraken@nagasaki-u.ac.jp .

N.T.H. and P.L.C. contributed equally to this work.

ABSTRACT

It is essential to continue the search for novel antimalarial drugs due to current spread of resistance against artemisinin by *Plasmodium falciparum* parasites. In this study, we developed *in silico* models to predict hemozoin inhibitors as a potential first-step screening for novel antimalarials. The *in vitro* colorimetric high throughput screening assay of hemozoin formation was used to identify hemozoin inhibitors from 9600 structurally diverse compounds. Physicochemical properties of positive hits and randomly selected compounds were extracted from ChemSpider database; they were used for developing prediction models to predict hemozoin inhibitors using two different approaches, i.e. traditional multivariate logistic regression, and Bayesian Modeling Average. Our results showed that a total of 224 positive hits exhibited the ability to inhibit the hemozoin formation with IC_{50} ranging from 3.1 µM to 199.5 µM. The "best" model according to traditional multivariate logistic regression included three variables: octanol-water partition coefficient, number of hydrogen bond donors, and number of atoms of hydrogen. Whereas, the "best" model according to Bayesian Modeling Average was octanol-water partition coefficient, number of hydrogen bond donors, and index of refraction. Both models had a good discriminatory power with the area under curve values were 0.736, and 0.781 for the traditional multivariate model, and the Bayesian Modeling Average model respectively. In conclusion, the prediction models can be a new, useful and cost-effective approach for the first screen of hemozoin inhibition based antimalarial drug discovery.

Hemozoin is a crystalline pigment product, which is synthesized by hemoparasites including *Plasmodium* species from the hemoglobin degradation process (1). Hemozoin formation is an adaptation of the parasite to be protected against toxic heme (2), which is released as a byproduct of hemoglobin degradation in the *Plasmodium* food vacuole. Within the infected red blood cells, the parasites digest hemoglobin as a main source of amino-acids for their growth and development (3). Due to the toxic effect of the released heme (4), it is imperative for *Plasmodium* to evolve an effective heme homeostasis mechanism, one of which is hemozoin formation (5).

The rapid spread of resistance to artemisinin-based combination therapies by *P. falciparum* parasites has been identified as a major global challenge in the fight against malaria (6, 7). Although the development of an effective malaria vaccine is the most effective control measure, there is still no available vaccine for preventing this disease (8). To date, only one malaria vaccine candidate has reached phase III clinical trials (9). It is essential to continue the search for novel antimalarial drugs, especially for malaria endemic countries. An ideal target is the blocking of the heme detoxification pathway of the parasite (10-13). Indeed, this mechanism is also one of the main targets of current antimalarial drugs like quinine, and has been the major target of several antimalarial screening projects. Unlike chloroquine resistance, resulting from mutation of membrane transport protein that effluxes chloroquine out of the food vacuole (1), quinine, although the reduced efficacy has been noticed recently, it still has strong antimalarial activity against chloroquine-resistant strains (14). This makes hemozoin inhibition a good target for novel antimalarial drug development.

Hemozoin formation is a physiochemical process that occurs in the presence of parasite proteins (15-18) and/or lipids (19, 20). Recently, the commercial lipophilic detergents including

Tween 20 and Nonidet P-40 (NP-40) have been identified as a surrogate substance to promote crystallization of heme under relevant conditions (21, 22). This artificial system is amenable for high-throughput hemozoin inhibition assays for screening novel antimalarials (23). However, it is still time consuming and requires expensive and specialized instruments and a laborious preparation. Therefore, the execution of *in silico* models or other machine learning models as Bayesian modelling are ideal for screening millions of chemical compounds to prioritize compounds for high-throughput screening (HTS) leading to valuable hit rates with fewer test compounds. Recently, Wicht *et al* showed that Bayesian models can be effective tools to predict hemozoin inhibitor compounds with high enrichment rates in comparison to conventional random screening (24). Making *in silico* models is not only valuable for future HTS, but it is also a good way to drive benefit from all available data, even inactives, from preceding screens. In this study, we developed a model to predict hemozoin inhibitors using physicochemical properties of chemical compounds.

MATERIALS AND METHODS

Materials. Hemin chloride (heme) and quinine were purchased from Sigma. "Core Library", which contains 9600 structurally diverse compounds, was from the Open Innovation Center for Drug Discovery, University of Tokyo (Tokyo, Japan). Detergent NP-40 served as a mediator for β-hematin formation due to its stability, low cost and low IC_{50} value, beside its similarity to the natural lipid particles of the parasite's vacuole (21, 25). Dimethyl sulfoxide (DMSO), from Wako Pure Chemicals, Osaka, Japan, was chosen as negative control because of its proven inability to inhibit the heme crystallization reaction (26).

High-throughput screening using anti-hemozoin assay. The assay was performed in a 384-well plate using quinine and DMSO as positive and negative controls, respectively (21, 27). Quinine and candidate compounds were dissolved in DMSO to achieve a final concentration of 220 µM. Using an automated dispenser (Multi-Dispensor EDR 384, BioTec, Japan), 5 µl of each compound was transferred to each well of the assay plates (see Fig. S1 in the supplemental material at https://www.researchgate.net/publication/309208397_Supplemental_material_Hig). Following the transfer of compounds, a Multidrop Combi dispenser (Thermo Fisher Scientific) was used to distribute 20 µl of hemin solution (10 mM heme in DMSO and 100 mM acetate buffer, pH = 4.8), as well as 10 µl of detergent NP-40 into each well of the plates. The assay mixture was incubated at 37°C for 250 minutes (25). Afterwards, pyridine solution was added to the mixture and shaken for 10 min. To dissolve the bubble, 10 µl acetone was added to each wells and the plate was finally transferred into a multi-plate reader to detect non-crystallized heme using the colorimetric method at 405/705 nm (27, 28).

Anti-hemozoin dose-response assay. Active compounds identified by the previously described high-throughput screening using anti-hemozoin assay were tested in dose-response assays. Quinine was also used as a positive control in each assay plate. Each compound's concentrations ranging from 0 µM to 208 µM were retested with the hemozoin inhibition assay in 384-well plates. The absorbance values of each compound measured at 405/750 nm were dependent on the difference in concentration of the compound. Data was analyzed to determine the half maximal inhibitory concentration (IC_{50}) for each compound, relying on a sigmoid dose-response curve fitted by GraphPad Prism software, version 5.00 (28).

Physicochemical properties of positive hits and representative sample of negative compounds. The average mass, octanol-water partition coefficient (Log P), distribution

coefficient (Log D), bio-concentration factor (BCF), adsorption coefficient (KOC), number of rule of five violations, number of hydrogen bond acceptors, number of hydrogen bond donors, freely rotating bonds, polar surface area, index of refraction, molar refractivity, molar volume, polarizability, flash point, boiling point, enthalpy of vaporization and number of atoms of chemical elements (such as bromine, carbon, chlorine, fluorine, hydrogen, nitrogen, oxygen and sulfur) of each of the positive hits and a sample of negative compounds were retrieved from ChemSpider (www.chemspider.com), as predicted by Advanced Chemistry Development (ACD/Laboratories) software (29).

Statistical Analysis. (i) Missing data analysis. We used complete case analysis, which delete compounds/ cases with missing data (i.e. physicochemical properties of a compound) so only complete compounds/ cases are left. The missing rates were variable from property to another. Therefore, they ranged from 0.2% to 3.5% of the compounds due to lack one of these physicochemical properties.

(ii) Univariate and Multivariate logistic regression. The outcome variable was the ability to inhibit the hemozoin formation of a compound, including two values: 1 if the compound can inhibit the hemozoin formation (exhibiting a typical sigmoid dose-response curve with IC_{50} < 200 μM), whereas 0 if the compound cannot. The predictor variables were physicochemical properties of a compound.

First of all, we performed univariate logistic regression to examine the association between physicochemical properties and the ability to inhibit hemozoin formation. Secondly, variables with p-values < 0.1 were submitted to multivariate analysis to find the independent predictors of inhibition of hemozoin formation. A significant level was set at *P* value < 0.05 in the multivariate regression.

(iii) Development and validation of the prediction models. The development and validation of the prediction models consist of following steps: (1) The original data was randomly divided into training and testing sets with the ratio 70:30 respectively. (2) The training data set was constructed to develop prediction models, using two approaches: one used the traditional approach, in which the univariate logistic regression was followed by the multivariate regression as described above; and another one used the Bayesian Modeling Averaging (BMA) approach to select the best prediction models. (3) The discriminatory powers of the best prediction models obtained from different approaches were compared on the basis of the area under the curve (AUC) from the receiver operating characteristic (ROC), and accuracy (30).

Basically, the purpose of BMA method is to search for the most parsimonious model (i.e., a model with the minimum number of explanatory variables and the maximum discriminatory power) (31). In brief, there are 2^k possible models (not including interaction models) can be constructed if there are k explanatory variables. Among 2^k models, the best models are suggested based on the Bayesian information criterion (BIC), in which a smaller BIC value indicates a better model. Therefore, unlike the tradition approach mentioned above, the BMA considers the "uncertainty" in the model selection process. Recently, BMA has been receiving more attention in prognosis model studies (32-34). All analyses were performed using R software version 3.2.2 (The R Foundation for Statistical Computing).

RESULTS

High-throughput screening (HTS) using the heme crystallization assay. Pyridine molecules formed coordinate bonds to free irons of non-crystallized heme molecules, and

produced a pyridine-heme complex with strong absorption at 405 nm (27). Robustness and reproducibility of the assay were improved by optimizing the concentration and volume of compounds, hemin, and detergent solutions. As a result, Z factors of all plates were higher than 0.5, which is an essential minimum value for validation of HTS assays. In other words, high degree of reproducibility and a large dynamic range were achieved for the assay (27).

A total of 9600 diversely selected compounds (the "core library"), were assigned randomly from more than 200,000 compounds in the chemical library of The Drug Discovery Initiative, Tokyo University (http://www.ddi.u-tokyo.ac.jp/en/#5), was used in HTS assay. Active compounds were identified as compounds with absorbance above three standard deviations of DMSO negative control. The absorbance values were described on 384-wells plate heat maps (see Fig. S1 in the supplemental material at https://www.researchgate.net/publication/309208397_Supplemental_material_Hig**).** Evident red color on plate heat maps represented correlative compounds, which were likely to strongly inhibit the crystallization of free heme. In total, 394 active compounds (4.1 % of 9,600 screened compounds) were identified by high throughput screening assay.

Dose-response assay for positive compounds. The 394 active compounds, resulting from HTS assay, were subsequently tested in dose-response assays to exclude false positives and to determine the half maximal inhibitory concentrations (IC_{50}). A positive hit was identified as an active compound exhibiting a typical sigmoid dose-response curve (see Fig. S2 in the supplemental material at https://www.researchgate.net/publication/309208397_Supplemental_material_Hig). False positives in which absorbance values could not generate typical sigmoid dose-response curves are likely due to compounds' colors or compounds aggregation.

Finally, 224 compounds out of 394 active compounds were shown to have positive hits (Fig. 1). Therefore, among 9,600 tested chemical compounds, both high throughput screening and dose-response assays identified 224 positive hits (resulting in a hit rate of 2.34%). Positive hits exhibited IC_{50}s from 3.1 µM to 199.5 µM, while 9 out of 224 positive hits exhibited IC_{50}s less than 5 µM.

Development of prediction models. The physical and chemical properties of all 224 positive hits as well as 199 negative compounds that were randomly selected from the original 9,600 compounds of the "core library" without anti-hemozoin activity, were extracted (Fig. 2). Then, 70% of the data (n=285, complete case analysis) was used to develop the "best" models using different approaches: traditional method (i.e. the univariate logistic regression was followed by the multivariate regression) and BMA method.

In traditional approach: Log P, KOC (pH 5.5 and pH 7.4), Log D (pH 5.5 and pH 7.4), index of refraction, molar refractivity, number of hydrogen bond donors, number of freely rotating bond donors, number of rule of five violations, density, surface tension, and number of atoms of hydrogen, oxygen, and nitrogen yielded p-values < 0.1 by univariate logistic regression analysis. The multivariate logistic regression of these properties showed that ability of positive hits to inhibit hemozoin formation was significantly correlated with Log P, number of hydrogen bond donors, and number of atoms of hydrogen with p -values < 0.05 (Table 1). The equation of the best multivariate model is represented as:

$$Logit\ (Probability) = -0.739 + 0.671 * Log\ P + 0.484 * N_1 - 0.099 * N_2$$

Where *Probability* represents the probability of anti-hemozoin activity of particular compound, while N_1 and N_2 stand for the number of hydrogen bond donors, and the number of atoms of hydrogen respectively.

In BMA approach: firstly, all variables yielded *P*-values < 0.1 by univariate logistic regression were submitted to BMA. Later, the BMA process suggested the five most parsimonious models on the basis of BIC values (Table 2). Among them, the "best" model included variables: Log P, number of hydrogen bond donors, and index of refraction which resulted the smallest BIC value. The equation of the best BMA model is represented as:

$$Logit\ (Probability) = -23.62 + \ 0.592 * Log\ P + 0.351 * N_1 + \ 1.322 * N_3$$

Where N_1 and N_3 stand for the number of hydrogen bond donors, and the index of refraction respectively.

Validation of prediction models. After we successfully developed the two "best" models (i.e. multivariate model and BMA model), they were validated using the 30% remains of the data (n=121). Figure 3 shows the AUC (left panel) and the accuracy (right panel) of these "best" models, which indicated that, the discriminatory power of the BMA model is better than that of the multivariate model. The AUC, however, were 0.736, and 0.781 for the multivariate model, and the BMA model respectively, it implies that both models have a good discriminatory power.

The multivariate model with a cut-off of 0.536 resulted an optimal sensitivity, specificity, and maximum accuracy at 65.6%, 77.2%, and 71.1% respectively. Whereas, the BMA model with a cut-off of 0.465 resulted an optimal sensitivity, specificity, and maximum accuracy at 79.7%, 66.7%, and 73.5% respectively.

DISCUSSION

The inhibition of hemozoin formation, proposed as the major mechanism of current antimalarials such as quinine and chloroquine (35), was the foundation of the research on novel antimalarials via high throughput screening assay. A total of 224 positive hits out of 9,600 library compounds exhibited the ability to inhibit the hemozoin formation with IC_{50}s ranging from 3.1 µM to 199.5 µM. Analysis of the physical and chemical properties of these positive hits showed positive correlation between Log P, index of refraction, number of hydrogen bond donors and capability to inhibit the hemozoin formation (Table 2).

The 2.34% hit rate fulfilled in our study is considerably higher than 0.42% in previous research by Sandlin *et al* (21). Compound concentration used in our assays was 220 µM, which is higher than concentrations in assays of Sandlin *et al*. Besides, 9600 compounds, used in this study that were assigned as a core chemical library with varieties of structural from more than 200,000 compounds, could be completely different from 38,400 compounds used in Sandlin *et al* or 5,000 compounds used in Wicht *et al* (24). Consequently, the difference in hit rates is likely due to the difference in the tested compound concentrations. The main advantage of our study over that was done by Wicht *et al* is that our study used two models, multivariate logistic regression and BMA, rather than BMA alone.

Octanol–water partition coefficient (Log P) of a compound expresses the tendency of a compound to partition between lipophilic phase and aqueous phase known as lipophilicity (36). Capability of compounds for hemozoin forming inhibition is probably related to compounds' lipophilicity as there is a very strong evidence supporting the lipid mediated formation theory of hemozoin (37-39). On examining the trophozoite stage of RBCs infected with *Plasmodium falciparum* by electron microscopy, Pisciotta *et al* found nanosphere lipid droplets containing

hemozoin crystals (40). Crystallization of β-haematin usually occur in a hydrophobic environment that is preferred for hydrogen bonds between the hydrophilic ferriprotoporphyrin IX's (Fe(III)PPIX) propionate linkage to be formed (41, 42). All these causes make the lipophilicity an important property of a compound enabling it to inhibit β-haematin or hemozoin formation. On cellular bases, lipophilic compounds can permeate through the lipid bilayer membrane of the food vacuole of *P. falciparum*, therefore can easily reach hemozoin crystals.

The index of refraction, also known as refractive index, is an optical property defined as the difference of velocity of light between the vacuum and the medium in which it propagates. In the Lorentz-Lorenz equation, refractive index is estimated using molar refraction, which is a sum of contributions of corresponding atoms and bonds (43). Hence, index of refraction related to polarizability, purity, density of organic compounds is applied to evaluate characteristics of the material (44). Moreover, presence and quantity of some heavy atoms and functional groups with high refractive index, such as sulfur (45), halogen elements (especially, bromine and iodine) (46), and phosphorus (47), play an important role in increasing the molar refraction. Therefore, presence of heavy atoms increases the index of refraction. However, our models could not detect an association of anti-hemozoin with any specific heavy atom, probably due to small sample size of each atom. Nevertheless, we were not able to find any research on the relationship between compound refractive indices and anti-hemozoin effect. Therefore, compounds with high refractive indices can be an interesting topic for antimalarial studies in the future.

The number of hydrogen bond donors plays an important role in β-hematin or haemozoin crystal inhibition through intramolecular formation of hydrogen bonds between neighboring complexes in the crystal (48). For instance, hydrogen bond donors of known antimalarial drugs, such as halofantrine and quinoline, form hydrogen bond bridge with β-hematin in parasite food

vacuoles (48, 49). Thus, these antimalarials are likely to inhibit hemozoin formation via hydrogen bond conformation (49). In summary, the number of hydrogen bond donors of compound candidates, positively related to anti-hemozoin capability of compounds, should be considered, on cellular base, for evaluation of various compounds' permeability and absorption based on Lipinski's rule (50).

In this study, we proposed two different approaches in the development of the prediction models for calculating the probability of inhibition of hemozoin formation by each compound. The AUC of both the multivariate model and the BMA model indicate that both models can be applied in a real setting (51). Interestingly, when testing against five well known antimalarial drugs including chloroquine, quinine, amodiaquine, halofantrine, and artemisinin, the BMA model accurately predicted all four well known anti-hemozoin drugs (*Probability:* chloroquine = 0.63, quinine = 0.60, amodiaquine = 0.88, halofantrine= 0.94) and one non anti-hemozoin drug (artemisinin = 0.12), while the multivariate model correctly predicted four drugs including chloroquine (0.58), amodiaquine (0.78), halofantrine (0.93) (probability > 0.536), and artemisinin (0.20), but wrongly predicted quinine (0.42) as a non-hemozoin inhibitor (probability < 0.536)..

The prediction models also have some advantages for the antimalarial design. Firstly, while other approaches such as development of analogs of existing agents or natural products mainly detect new antimalarials by the chemical modifications of known compounds (52), new antimalarial compounds can be discovered by the prediction equation based on the well-known metabolic target. Thus, the models help researchers to find out the good chemical groups for synthetic compounds. Secondly, the expensive equipment and specialized labwares are not essential in these prediction models. Therefore, millions of library compounds can be screened

in-silico by using the models. Thirdly, the relationship between the properties of compounds and anti-hemozoin activity is also interpreted from the models. It can be the first clue for understanding the mechanism of action of antimalarials. In addition, we proved that BMA is likely a good approach in the development of prediction models because it considers the "uncertainty" in model selections. Hence, the habit of building the only-right model in traditional approach should be compared with other approaches which consider "uncertainty" in model selections (e.g. BMA) in similar studies to this one.

Besides the benefits of high throughput screening assay already mentioned, the study has several limitations. First, in the anti-hemozoin assay, we did not remove all the soluble contents before dissolving the non-crystallized heme by adding pyridine solution (35). Lack this step probably resulted in the false positives due to compounds' color and/or aggregation. These were eliminated in the second step using dose-response assay. Secondly, the interaction between the positive hits and intra-parasitic condition was not fully evaluated in this paper, although the assay was performed under conditions that closely mimic the physiological environment in the parasite food vacuole. It is known that only a small fraction of hemozoin inhibitors possesses an antimalarial activity *in vitro*. Our ongoing experiments revealed a total of 23 positive hit compounds and two negative hit compounds exhibited antimalarial activity with IC_{50} value less than 10 μM. Among them, four compounds of positive hits showed IC_{50} below 1 μM. However, using anti-hemozoin as a HTS, we could lower the *in vitro* antimalarial assay workload approximately 40 times. The last limitation is that the prediction models have not been validated yet in external samples.

In conclusion, the *in vitro* high-throughput hemozoin formation assay was performed with a high degree of reproducibility and robustness. A total 224 true positive hits were

identified from the "core library" with a hit rate of 2.34%. The prediction models based on physicochemical parameters represent a new, useful and cost-effective approach for antimalarial drug discovery in developing countries. Moreover, the physicochemical properties, namely: log P, index of refraction, and the number of hydrogen bond donors should be investigated further in order to find out their effects on the anti-hemozoin activity of compounds.

Acknowledgements: This work was partly supported by Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science and Technology, Japan. The funder had no role in study design, data collection and analysis, preparation of the manuscript, or decision to publish.

REFERENCES

1. **Hänscheid T, Egan TJ, Grobusch MP.** 2007. Haemozoin: from melatonin pigment to drug target, diagnostic tool, and immune modulator. Lancet Infect Dis **7:**675-685.
2. **Ihekwereme CP, Esimone CO, Nwanegbo EC.** 2014. Hemozoin inhibition and control of clinical malaria. Adv Pharmacol Sci **2014:**984150.
3. **Zhang J, Krugliak M, Ginsburg H.** 1999. The fate of ferriprotorphyrin IX in malaria infected erythrocytes in conjunction with the mode of action of antimalarial drugs. Mol Biochem Parasitol **99:**129-141.
4. **Fitch CD, Chevli R, Kanjananggulpan P, Dutta P, Chevli K, Chou AC.** 1983. Intracellular ferriprotoporphyrin IX is a lytic agent. Blood **62:**1165-1168.
5. **Moore LR, Fujioka H, Williams PS, Chalmers JJ, Grimberg B, Zimmerman PA, Zborowski M.** 2006. Hemoglobin degradation in malaria-infected erythrocytes determined from live cell magnetophoresis. FASEB J **20:**747-749.
6. **Dondorp AM, Yeung S, White L, Nguon C, Day NP, Socheat D, von Seidlein L.** 2010. Artemisinin resistance: current status and scenarios for containment. Nat Rev Microbiol **8:**272-280.
7. **Dondorp AM, Ringwald P.** 2013. Artemisinin resistance is a clear and present danger. Trends Parasitol **29:**359-360.

8. **Targett GA, Moorthy VS, Brown GV.** 2013. Malaria vaccine research and development: the role of the WHO MALVAC committee. Malar J **12:**362.

9. **Agnandji ST, Lell B, Fernandes JF, Abossolo BP, Methogo BG, Kabwende AL, Adegnika AA, Mordmüller B, Issifou S, Kremsner PG.** 2012. A phase 3 trial of RTS, S/AS01 malaria vaccine in African infants. The New England journal of medicine **367:**2284-2295.

10. **Egan TJ, Hunter R, Kaschula CH, Marques HM, Misplon A, Walden J.** 2000. Structure-function relationships in aminoquinolines: effect of amino and chloro groups on quinoline-hematin complex formation, inhibition of beta-hematin formation, and antiplasmodial activity. J Med Chem **43:**283-291.

11. **Adams PA, Berman PA, Egan TJ, Marsh PJ, Silver J.** 1996. The iron environment in heme and heme-antimalarial complexes of pharmacological interest. J Inorg Biochem **63:**69-77.

12. **Sullivan DJ.** 2002. Theories on malarial pigment formation and quinoline action. Int J Parasitol **32:**1645-1653.

13. **Ridley RG, Dorn A, Vippagunta SR, Vennerstrom JL.** 1997. Haematin (haem) polymerization and its inhibition by quinoline antimalarials. Ann Trop Med Parasitol **91:**559-566.

14. **Dinio T, Gorka AP, McGinniss A, Roepe PD, Morgan JB.** 2012. Investigating the activity of quinine analogues versus chloroquine resistant Plasmodium falciparum. Bioorg Med Chem **20:**3292-3297.

15. **Sullivan DJ, Jr., Gluzman IY, Goldberg DE.** 1996. Plasmodium hemozoin formation mediated by histidine-rich proteins. Science **271:**219-222.

16. **Huy NT, Serada S, Trang DT, Takano R, Kondo Y, Kanaori K, Tajima K, Hara S, Kamei K.** 2003. Neutralization of toxic heme by Plasmodium falciparum histidine-rich protein 2. J Biochem (Tokyo) **133:**693-698.

17. **Jani D, Nagarkatti R, Beatty W, Angel R, Slebodnick C, Andersen J, Kumar S, Rathore D.** 2008. HDP-a novel heme detoxification protein from the malaria parasite. PLoS Pathog **4:**e1000053.

18. **Nakatani K, Ishikawa H, Aono S, Mizutani Y.** 2014. Identification of essential histidine residues involved in heme binding and Hemozoin formation in heme detoxification protein from Plasmodium falciparum. Sci Rep **4:**6137.

19. **Dorn A, Stoffel R, Matile H, Bubendorf A, Ridley RG.** 1995. Malarial haemozoin/beta-haematin supports haem polymerization in the absence of protein. Nature **374:**269-271.

20. **Tripathi AK, Gupta A, Garg SK, Tekwani BL.** 2001. In vitro beta-hematin formation assays with plasma of mice infected with Plasmodium yoelii and other parasite preparations: comparative inhibition with quinoline and endoperoxide antimalarials. Life Sci **69:**2725-2733.

21. **Sandlin RD, Carter MD, Lee PJ, Auschwitz JM, Leed SE, Johnson JD, Wright DW.** 2011. Use of the NP-40 Detergent-Mediated Assay in Discovery of Inhibitors of β-Hematin Crystallization. Antimicrobial Agents and Chemotherapy **55:**3363-3369.

22. **Huy NT, Uyen DT, Maeda A, Oida T, Harada S, Kamei K.** 2007. Simple colorimetric inhibition assay of heme crystallization for high-throughput screening of antimalarial compounds. Antimicrobial agents and chemotherapy **51:**350-353.

23. **Sandlin RD, Carter MD, Lee PJ, Auschwitz JM, Leed SE, Johnson JD, Wright DW.** 2011. Use of the NP-40 detergent-mediated assay in discovery of inhibitors of beta-hematin crystallization. Antimicrob Agents Chemother **55:**3363-3369.

24. **Wicht KJ, Combrinck JM, Smith PJ, Egan TJ.** 2015. Bayesian models trained with HTS data for predicting β-haematin inhibition and in vitro antimalarial activity. Bioorganic & medicinal chemistry **23:**5210-5217.

25. **Carter MD, Phelan VV, Sandlin RD, Bachmann BO, Wright DW.** 2010. Lipophilic mediated assays for β-hematin inhibitors. Combinatorial chemistry & high throughput screening **13:**285.

26. **Kurosawa Y, Dorn A, Kitsuji-Shirane M, Shimada H, Satoh T, Matile H, Hofheinz W, Masciadri R, Kansy M, Ridley RG.** 2000. Hematin Polymerization Assay as a High-Throughput Screen for Identification of New Antimalarial Pharmacophores. Antimicrobial Agents and Chemotherapy **44:**2638-2644.

27. **Rush MA, Baniecki ML, Mazitschek R, Cortese JF, Wiegand R, Clardy J, Wirth DF.** 2009. Colorimetric High-Throughput Screen for Detection of Heme Crystallization Inhibitors. Antimicrobial Agents and Chemotherapy **53:**2564-2568.

28. **Vargas S, Ndjoko Ioset K, Hay A-E, Ioset J-R, Wittlin S, Hostettmann K.** 2011. Screening medicinal plants for the detection of novel antimalarial products applying the inhibition of β-hematin formation. Journal of pharmaceutical and biomedical analysis **56:**880-886.

29. **Huy NT, Shima Y, Maeda A, Men TT, Hirayama K, Hirase A, Miyazawa A, Kamei K.** 2013. Phospholipid Membrane-Mediated Hemozoin Formation: The Effects of

Physical Properties and Evidence of Membrane Surrounding Hemozoin. PloS one **8:**e70025.

30. **Jin H, Ling CX.** 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering **17:**299-310.

31. **Jennifer A. Hoeting DM, Adrian E. Raftery, Chris T. Volinsky.** 1999. Bayesian Model Averaging: A Tutorial. Statistical Science **14:**382-417.

32. **Wang D, Zhang W, Bakhai A.** 2004. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. Stat Med **23:**3451-3467.

33. **Binh TQ, Thu NTT, Phuong PT, Nhung BT, Nhung TTH.** 2015. CDKN2A-rs10811661 polymorphism, waist-hip ratio, systolic blood pressure, and dyslipidemia are the independent risk factors for prediabetes in a Vietnamese population. BMC Genet **16**.

34. **Tran TS, Hirst JE, Do MAT, Morris JM, Jeffery HE.** 2013. Early Prediction of Gestational Diabetes Mellitus in Vietnam: Clinical impact of currently recommended diagnostic criteria. Diabetes Care **36:**618-624.

35. **Tekwani BL, Walker LA.** 2005. Targeting the hemozoin synthesis pathway for new antimalarial drug discovery: technologies for in vitro beta-hematin formation assay. Combinatorial chemistry & high throughput screening **8:**63-79.

36. **Tetko IV, Tanchuk VY, Villa AE.** 2001. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. Journal of chemical information and computer sciences **41:**1407-1421.

37. **Bendrat K, Berger BJ, Cerami A.** 1995. Haem polymerization in malaria. Nature **378:**138-139.

462 38. **Dorn A, Vippagunta SR, Matile H, Bubendorf A, Vennerstrom JL, Ridley RG.** 1998.
463 A comparison and analysis of several ways to promote haematin (haem) polymerisation
464 and an assessment of its initiation in vitro. Biochemical pharmacology **55:**737-747.
465 39. **Fitch CD, Cai G-z, Chen Y-F, Shoemaker JD.** 1999. Involvement of lipids in
466 ferriprotoporphyrin IX polymerization in malaria. Biochimica et Biophysica Acta (BBA)-
467 Molecular Basis of Disease **1454:**31-37.
468 40. **Pisciotta JM, Coppens I, Tripathi AK, Scholl PF, Shuman J, Bajad S, Shulaev V,**
469 **Sullivan DJ.** 2007. The role of neutral lipid nanospheres in Plasmodium falciparum haem
470 crystallization. Biochemical Journal **402:**197-204.
471 41. **Egan TJ, Chen JYJ, de Villiers KA, Mabotha TE, Naidoo KJ, Ncokazi KK,**
472 **Langford SJ, McNaughton D, Pandiancherri S, Wood BR.** 2006. Haemozoin (β-
473 haematin) biomineralization occurs by self-assembly near the lipid/water interface. FEBS
474 letters **580:**5105-5110.
475 42. **Huy NT, Maeda A, Uyen DT, Trang DTX, Sasai M, Shiono T, Oida T, Harada S,**
476 **Kamei K.** 2007. Alcohols induce beta-hematin formation via the dissociation of
477 aggregated heme and reduction in interfacial tension of the solution. Acta tropica
478 **101:**130-138.
479 43. **Gharagheizi F, Ilani-Kashkouli P, Kamari A, Mohammadi AH, Ramjugernath D.**
480 2014. Group Contribution Model for the Prediction of Refractive Indices of Organic
481 Compounds, p 1930–1943. *In* Journal of Chemical & Engineering Data.
482 44. **Katritzky AR, Sild S, Karelson M.** 1998. General quantitative structure-property
483 relationship treatment of the refractive index of organic compounds. Journal of chemical
484 information and computer sciences **38:**840-844.

45. **Liu J-g, Nakamura Y, Shibasaki Y, Ando S, Ueda M.** 2007. High refractive index polyimides derived from 2, 7-bis (4-aminophenylenesulfanyl) thianthrene and aromatic dianhydrides. Macromolecules **40:**4614-4620.

46. **Gaudiana RA, Minns RA, Rogers HG.** 1992. High refractive index polymers. US.

47. **Olshavsky M, Allcock HR.** 1997. Polyphosphazenes with high refractive indices: optical dispersion and molar refractivity. Macromolecules **30:**4179-4183.

48. **de Villiers KA, Marques HM, Egan TJ.** 2008. The crystal structure of halofantrine–ferriprotoporphyrin IX and the mechanism of action of arylmethanol antimalarials. Journal of inorganic biochemistry **102:**1660-1667.

49. **Buller R, Peterson ML, Almarsson Ö, Leiserowitz L.** 2002. Quinoline binding site on malaria pigment crystal: a rational pathway for antimalaria drug design. Crystal growth & design **2:**553-562.

50. **Lipinski CA, Lombardo F, Dominy BW, Feeney PJ.** 2012. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews **64:**4-17.

51. **Zou KH, O'Malley AJ, Mauri L.** 2007. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation **115:**654-657.

52. **Rosenthal PJ.** 2003. Antimalarial drug discovery: old and new approaches. Journal of Experimental Biology **206:**3735-3744.

FIGURE LEGENDS

FIG 1 Scheme for building a prediction model of anti-hemozoin compounds. Positive hits were firstly identified from 9600 structurally diverse compounds by high-throughput screening (HTS) and dose-response assay of hemozoin formation. Secondly, physical properties of 224 true positive hits and 199 random negative compounds were extracted using the ChemSpider software. Thirdly, prediction models were built by traditional approach vs. Bayesian approach using these physical properties.

FIG 2 Anti-hemozoin HTS of 9,600 diverse compounds. Dots represents % hemozoin inhibition of 9600 compounds including 224 true positive hits (closed dots above cut-off line), 170 false positive compounds (open dots above cut-off line), and negatives (dots under cut-off line). Cut-off value was determined as average absorbance value of negative DMSO control plus 3 standard deviations by each HTS anti-hemozoin assay. The true positive hits were identified by subsequent dose-response assay.

FIG 3 The discriminatory powers comparison of the best multivariate model and the best BMA model on the basis of AUC (left panel) and accuracy (right panel). The best multi variate model consists of three variables: Log P, the number of hydrogen bond donors, and the number of atoms of hydrogen. Whereas, the best BMA model consists of three variables: Log P, the number of hydrogen bond donors, and the index of refraction. The discriminatory power of the BMA model is better than that of the multivariate model in term of AUC and accuracy.

TABLE 1 Univariate and multivariate analyses of positive hit versus negative hit compounds.

Predictors	Univariate analysis		Multivariate analysis	
	OR (95%CI)	**p value**	**Adjusted OR (95%CI)**	***P* value**
Log P	1.54 (1.29-1.83)	<0.0001	2.04 (1.27-3.27)	0.0028
BCF_ pH 5.5	1.00 (1.00-1.00)	0.2181	//	//
BCF_pH7.4	1.00 (1.00-1.00)	0.185	//	//
KOC_pH5.5	1.00 (1.00-1.00)	0.052	1.00 (1.00-1.00)	0.9400
KOC_pH7.4	1.00 (1.00-1.00)	0.041	1.00 (1.00-1.00)	0.4344
LogD_pH5.5	1.34 (1.20-1.50)	<0.0001	1.05 (0.69-1.62)	0.7939
LogD_pH7.4	1.34 (1.18-1.52)	<0.0001	0.88 (0.60-1.30)	0.5458
Average Mass (Da)	1.00 (1.00-1.00)	0.251	//	//
Density (g/cm3)	17.10 (3.59-82.10)	0.0004	0.35 (0.01-18.5)	0.6088
Index of refraction*	3.72 (2.38-5.81)	<0.0001	1.55 (0.64-3.76)	0.3304
Molar refractivity (cm^3)	1.01 (1.00-1.02)	0.0686	1.0. (0.99-1.06)	0.0976
Molar volume (cm^3)	1.00 (0.99-1.00)	0.4041	//	//
Mono isotopic mass (Da)	1.00 (1.00-1.00)	0.3383	//	//
No Freely rotating bonds	0.92 (0.82-1.01)	0.0933	0.95 (0.79-1.16)	0.6712
No H bond acceptors	0.98 (0.86-1.11)	0.7777	//	//
No H bond donors	1.40 (1.13-1.73)	0.0017	1.38 (1.03-1.87)	0.0308
No of rule of 5 violations	4.00 (1.45-10.98)	0.0071	3.42 (0.72-16.31)	0.1218
Number of Br	1.79e+7 (0-Inf)	0.9855	//	//
Number of C	1.02 (0.96-1.07)	0.4763	//	//
Number of Cl	1.37 (0.79-2.37)	0.2517	//	//
Number of F	0.70 (0.46-1.06)	0.1000	//	//
Number of H	0.94 (0.91-0.98)	0.0105	0.87 (0.77-0.98)	0.0293

Number of N	1.25 (1.06-1.46)	0.0066	1.09 (0.82-1.46)	0.5150
Number of O	0.81 (0.70-0.93)	0.0049	0.81 (0.59-1.10)	0.1795
Number of S	1.13 (0.82-1.57)	0.4311		
Polar surface area (Å2)	1.00 (0.99-1.01)	0.2108	//	//
Polarizability ($\times 10^{24}$ cm^3)	1.02 (0.99-1.05)	0.1016	//	//
Surface tension (dyne/cm)	1.04 (1.02-1.06)	<0.0001	1.01 (0.96-1.06)	0.4724

LogP, octanol-water partition coefficient; LogD, distribution coefficient; BCF, bio-concentration factor; KOC, adsorption coefficient; OR, odds ratio
Significant P values (<0.05) in multivariate analysis were underlined.
// These variables were not included in multivariate analysis because these variables had p value >= 0.1 in univariate analysis.

*The original scale has been multiplied by 10

TABLE 2 The five most parsimonious models selected by Bayesian Model Average (BMA)

approach

Model	Explanatory variables	Coefficient	P value	BIC**	Posterior probability***
1	Log P	0.5924	3.05e-09	-1278	0.318
	No H bond donors	0.3514	0.00495		
	Index of refraction*	1.322	0.00215		
	Intercept = -23.6		0.00192		
2	Log P	0.5328	1.26e-10	-1276	0.118
	Index of refraction*	1.487	1.79e-05		
	Intercept = -25.5		0.00098		
3	Number of H	-0.0519	1.18e-09	-1276	0.117
	Log P	0.6506	5.37e-05		
	No H bond donors	0.3600	0.00104		
	Index of refraction*	1.149	0.01276		
	Intercept = -20.05				
4	Number of O	-0.1374	9.2e-09	-1275	0.066
	Log P	0.5774	0.008566		
	No H bond donors	0.3778	0.001341		
	Index of refraction*	1.247	0.000925		
	Intercept =-22.0				
5	Log P	0.6267	2.83e-10	-1275	0.054
	No H bond donors	0.3862	1.16e-05		
	No Freely rotating bonds	-0.0946	0.000514		
	Index of refraction*	1.206			
	Intercept = -21.43				

Through the BMA process, 18 models were selected and 5 best models were presented. The cumulative posterior probability is equal to 0.6727

*The original scale has been multiplied by 10

**BIC stands for Bayesian Information Criteria. BIC smallest suggested the model with maximum parsimony (i.e. minimum explanatory variables and maximum discrimination power)

***Posterior probability is the probability of a model being a "correct" model in BMA process