Bioinformatics Advances, 2023, **00**, vbad112
https://doi.org/10.1093/bioadv/vbad112
Advance Access Publication Date: 21 August 2023
Original Article

OXFORD

Systems biology

QuTIE: quantum optimization for target identification by enzymes

Hoang M. Ngo[1], My T. Thai[1,*], Tamer Kahveci[1,*]

[1]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, United States

*Corresponding authors. Department of Computer and Information Science and Engineering, University of Florida, Newell Dr, Gainesville, FL 32611, United States. E-mails: mythai@cise.ufl.edu (M.T.T.) and tamer@cise.ufl.edu (T.K.)

Associate Editor: Thomas Lengauer

Abstract

Summary: Target identification by enzymes (TIE) problem aims to identify the set of enzymes in a given metabolic network, such that their inhibition eliminates a given set of target compounds associated with a disease while incurring minimum damage to the rest of the compounds. This is a NP-hard problem, and thus optimal solutions using classical computers fail to scale to large metabolic networks. In this article, we develop the first quantum optimization solution, called *QuTIE* (quantum optimization for target identification by enzymes), to this NP-hard problem. We do that by developing an equivalent formulation of the TIE problem in quadratic unconstrained binary optimization form. We then map it to a logical graph, and embed the logical graph on a quantum hardware graph. Our experimental results on 27 metabolic networks from *Escherichia coli*, *Homo sapiens*, and *Mus musculus* show that QuTIE yields solutions that are optimal or almost optimal. Our experiments also demonstrate that QuTIE can successfully identify enzyme targets already verified in wet-lab experiments for 14 major disease classes.

Availability and implementation: Code and sample data are available at: https://github.com/ngominhhoang/Quantum-Target-Identification-by-Enzymes.

1 Introduction

Enzymes catalyze reactions that operate on and transform a set of compounds. The compounds that are input to a reaction are called substrates, and those that are produced once that reaction completes are called products. Reactions take place in a complex network topology, where the products of a set of reactions can be substrates for another set of reactions. This organization of reactions is also called the metabolic network. Over- or underproduction of certain compounds in metabolism can lead to serious disorders. For instance, the abundance of dopamine is linked with the development and severity of the Alzheimer's disease (Pritchard *et al.* 2009, Martorana *et al.* 2014). Similarly, Hunter syndrome is caused by the metabolism's inability to breakdown sugar molecules (D'Avanzo *et al.* 2020), and the malfunction of enzyme phosphatidic acid phosphatase leads to the overexpression of lipin, causing obesity (Carman and Han 2009).

One way to address such compound-based disorders is to alter and regulate the production of such compounds by targeting a small subset of enzymes with drugs (Robertson, 2007, Li *et al.* 2020, Berillo *et al.* 2021), as enzymes are potential drug targets when other drug targets, such as cell surface receptors, DNA, and transporters are not possible to target, or they do not yield the desired impact on the compound regulation (Robertson 2005). Drugs that target a specific enzyme inhibit that enzyme and slow down or stop the reactions catalyzed by that enzyme, and thus regulate the abundance of a subset of compounds by stopping/slowing down their production produced downstream of those reactions (Terentis *et al.* 2010, Cooney, 2017).

Although it is possible to regulate the metabolic network by targeting enzymes, unintended consequences can happen as a result of this process for various reasons. For instance, the enzymes inhibited by the underlying drug may be responsible from the catalysis of multiple reactions. Some of these reactions can produce compounds that lead to the underlying disorder (i.e. intended targets), while others consume/produce different compounds that are unrelated to the disorder (i.e. unintended targets). While reducing the abundance of compounds in the first category is desirable, doing that for the second category may lead to other problems, called side-effects (Sridhar *et al.* 2008, Mizutani *et al.* 2012).

One of the fundamental goals in drug development is to obtain a balance between the two potentially conflicting outcomes, namely efficacy and toxicity of the drug (Shankarappa *et al.* 2014). Efficacy measures how well the desired outcome (such as lowering the blood pressure if that is the goal of the drug) is achieved, while toxicity measures the damage inflicted on the organism (Cohen *et al.* 2010, Riley and Kohut, 2010). In order to formulate these concepts mathematically, we call the compounds that are intended to be inhibited (i.e. the compounds whose overproduction causes the underlying disorder) target compounds, and the remaining ones nontarget compounds. A given enzyme-binding drug limits the production of a set of compounds, some of which are target, while others are not (Copeland *et al.* 2007). One way to formulate the toxicity of a drug is in terms of the

Received: 15 May 2023; **Revised:** 26 June 2023; **Editorial Decision:** 1 August 2023; **Accepted:** 18 August 2023
Published by Oxford University Press 2023.

number of nontarget compounds that are inhibited (this number is also called damage) (Choi, 2008, Sridhar *et al.* 2008).

Following from the definitions above, given a metabolic network, including a set of enzymes, reactions, compounds, and relations amongst them, the target identification by enzymes (TIE) problem aims to identify the set of enzymes, such that their inhibition eliminates a given set of target compounds, while incurring minimum damage. This is an NP-hard problem (Song *et al.* 2011) and there are several approaches that address the TIE problem. However, these solutions do not scale well due to exponential complexity of the TIE problem (see Supplementary Material S1—SM 1 for details).

Recently, quantum computing has shown its supremacy over classical computers in some tasks that are intractable using classical computers, such as finding prime factors of large integer (Shor 1999). While quantum computing is at a very early stage of development (Preskill 2018), quantum-inspired methods have already been developed in a wide range of fields, such as machine learning (Jerbi *et al.* 2021) and optimization (Guillaume *et al.* 2022). The fundamental limitation of quantum computing currently is that their capacity in qubits (i.e. quantum bit that is the quantum analog of the bit classical computers) is limited. One of the most outstanding paradigms that overcomes this limitation for quantum computing is quantum annealing (QA). This paradigm focuses on solving optimization problems by utilizing quantum fluctuation. QA scales to significantly larger number of qubits than other types of quantum computing. This characteristic enables QA to solve large optimization tasks in bioinformatics, such as designing peptides (Mulligan *et al.* 2019), RNA folding (Fox *et al.* 2022), and DNA sequence assembly (Nałęcz-Charkiewicz and Nowak 2022).

QA has three major steps. The first step formulates the optimization problem in quadratic unconstrained binary optimization (QUBO) form. It maps the resulting QUBO on a graph, called logical graph, and then maps the logical graph into quantum processing unit (QPU) whose topology is represented by another graph, called hardware graph. The final step assigns appropriate parameters to the hardware graph, and runs QA to find candidate solutions for the optimization problem (see Supplementary Material S1—SM 2 for details).

1.1 Contributions

In this article, we consider the TIE problem, and develop the first quantum optimization solution, called QuTIE (quantum optimization for target identification by enzymes), to this NP-hard problem. We formulate the TIE problem in the QUBO form, and map the enzymes and metabolic reactions to nodes and edges in the logical graph. We utilize QA to find optimal solutions for the TIE problem by mapping the logical graph on the hardware graph. We implement and test our solution on the quantum hybrid framework, the largest quantum annealing system available. We compare our method against four methods operating on classic computers: the exact method (OPMET), the IP method, the heuristic method (double iterative), and the simulated annealing (SA) method. Our results on 27 datasets from *Escherichia coli*, *Homo sapiens*, and *Mus musculus* metabolic network collected from *KEGG* database show that QuTIE yields solutions that are optimal, or close to optimal. Our method outperforms the existing methods for large datasets in which the exact method cannot run. Our experiments on the biosynthesis of amino acids network of *H.sapiens* demonstrate that QuTIE can successfully identify enzyme targets already verified in wet-lab experiments for 14 major disease classes. In addition to solving the NP-hard drug target identification, using quantum optimization, this article lays the background and opens the door for formulating and solving high-complexity problems studying biological networks using quantum computing.

2 Methods

Here, we first define the TIE problem. We then describe the objective function for the TIE problem in the QUBO form. There are four parts in the objective function, namely, damage scoring function, target penalty function, reaction inference penalty function, and compound inference penalty function.

2.1 Formal definition

Consider a set of enzymes E, a set of reactions R, and a set of compounds C. Metabolic network shows the relationship between the entities in these three sets. Specifically, enzymes catalyze reactions. Each reaction consumes a set of compounds, and produces another set of compounds. We represent these relationships in a metabolic network as a directed graph $G = (V_G, E_G)$. In this tuple representation, the first term is the union of three mutually exclusive sets of nodes $V_G = E \cup R \cup C$. Each node in V_G corresponds to either an enzyme, reaction, or compound. The second term, E_G denotes the set of directed edges among those nodes. Consider two nodes $u, v \in V_G$. Each directed edge (u, v) from node u to v represents one of the three possible types of relations among the nodes as follows:

1) The enzyme corresponding to node $u \in E$ catalyzes the reaction corresponding to node $v \in R$.
2) The compound corresponding to node $u \in C$ is a substrate, consumed by the reaction denoted with $v \in R$.
3) The reaction corresponding to node $u \in R$ produces the compound denoted with $v \in C$.

Notice that, the above mathematical model expresses a metabolic network as a closed system, before the introduction of enzyme binding drug molecules. Following from these observations, we list the inhibition conditions for each node in $u \in G$ depending on what that node represents as follows:

- **Condition 1:** An enzyme is inhibited when an enzyme binding drug molecule binds to it.
- **Condition 2:** A reaction denoted by node $r \in R$ is inhibited if at least one of the two conditions is satisfied: (i) If there is an input compound, denoted by $c \in C$, consumed by that reaction is inhibited, or (ii) if at least one of the enzymes, denoted by $e \in E$, which catalyzes that reaction is inhibited.
- **Condition 3:** A compound denoted by $c \in C$ is inhibited if all the reactions denoted by $r \in R$ which produce that compound are inhibited.

Based on the three conditions above, given a set of compounds, called target set $C_{\text{target}} \subseteq C$, it is possible to find a subset of enzymes in E whose inhibition eliminates the production of all compounds in the target set. One can prove the statement above by inhibiting all the enzymes (i.e. all nodes in E), thus stopping the production of all the compounds, including those in C_{target}. This, however, is more than what is needed, as it also eliminates the compounds in $C - C_{\text{target}}$, as

well. This is undesirable for the compounds in $C - C_{\text{target}}$ are needed for the healthy metabolism. We measure the number of compounds in $C - C_{\text{target}}$ whose production stops as a result of inhibition of a subset of enzymes as the damage of inhibiting that subset of enzymes (see the example in Supplementary Fig. S1). We desire to inhibit all the compounds in the target set (i.e. maximum efficacy) with minimum damage (i.e. minimum toxicity). Let us denote the set of compounds whose production stops as a result of inhibiting a set of enzymes $E' \subseteq E$ as $C_{E'} \subseteq C$, and the damage to the metabolic network G as the cardinality of the set $C_{E'} - C_{\text{target}}$. We formally define the TIE problem as:

Definition 1 Consider a metabolic network consisting of a set of enzymes, a set of reactions, and a set of compounds. Let us denote the set of nodes corresponding to these three sets with E, R, and C, respectively. Let us denote this network with $G = (V_G, E_G)$, where $V_G = E \cup R \cup C$. Given a target set of compounds $C_{\text{target}} \subseteq C$. TIE problem seeks for a set of enzymes $E^\star$ such that:

$$E^\star = \text{argmin}\{|C_{E'}| : E' \subseteq E \text{ AND } C_{\text{target}} \subseteq C_{E'}\} \quad (1)$$

2.2 QUBO construction for TIE problem

Before constructing QUBO for the TIE problem, we develop a Boolean model for the TIE problem. Given the metabolic network $G = (V_G, E_G)$ with $V_G = E \cup R \cup C$, we denote the state of each node $u \in V_G$ in the metabolic network with a binary variable x_u such that:

$$x_u = \begin{cases} 0 & \text{if } u \text{ isnotinhibited.} \\ 1 & \text{if } u \text{ isinhibited.} \end{cases} \quad (2)$$

To satisfy the constraints of the TIE problem, all compounds in the target set C_{target} need to be inhibited. We express this constraint as:

$$\prod_{c \in C_{\text{target}}} x_c = 1 \quad (3)$$

Next, we present inhibition conditions as follows:

- **Condition 1:** The state of an enzyme $e \in E$ only depends on x_e.
- **Condition 2:** Consider a node in the metabolic network corresponding to a reaction, $r \in R$. Let us define the set of nodes in its immediate upstream with the set $N(r)$. Recall that a node $v \in N(r)$ if one of the two criteria is satisfied: (i) v corresponds to an enzyme which catalyzes the reaction corresponding to node r, and (ii) v corresponds to a compound which is consumed by the reaction corresponding to node r. Thus, we have $v \in (E \cup C)$. If any of the nodes in $N(r)$ are inhibited, that implies r is also inhibited. As a result, the state of the reaction r is valid if it satisfies:

$$x_r = 1 - \prod_{v \in N(r)} (1 - x_v) \quad (4)$$

- **Condition 3:** Consider a node in the metabolic network corresponding to a compound, $c \in C$. Let us define the set of nodes in its immediate upstream with set $N(c)$. Recall that each node $v \in N(c)$ corresponds to a reaction that produces the compound denoted by node c, thus $v \in R$. If all of the nodes in $N(c)$ are inhibited, then c is also inhibited. As a result, a state of a compound c is valid if it satisfies:

$$x_c = \prod_{v \in N(c)} x_v \quad (5)$$

We acknowledge the presence of a scenario where a group of alternative enzymes catalyzes a reaction, and the inhibition of the reaction occurs only when all enzymes in the group are inhibited. In such cases, we can reformulate (4) in a similar manner to (5).

We consider an assignment of values to the set of variables $\{x_u | u \in E \cup R \cup C\}$ is valid if all x_u satisfy constraints (3), (4), and (5).

The critical challenge we need to address to solve the TIE problem by QA is to represent the TIE problem as a QUBO function. Thus, our goal is to design an energy function H for the TIE problem in form of QUBO. H takes a set of binary variables as input (the input binary set also includes auxiliary binary variables that we discuss later). As H is unconstrained, we need to discriminate invalid and valid assignments of values to the input variables of this function. Furthermore, the function H must return values corresponding to the damage of input assignment. To sum up, in order to model the TIE problem, function H must follow two principles:

- The value of H for a valid assignment must be lower than that for every invalid assignment.
- The value of H for a valid assignment must be equal to the damage produced by that assignment.

Minimizing the value of function H which follows these two principles is equivalent to finding a valid assignment with minimum damage for the underlying TIE problem.

Based on two above principles, we construct the energy function H as a combination of four quadratic functions, namely, Damage Scoring, Target Penalty, Reaction Inference Penalty, and Compound Inference Penalty. Damage Scoring function measures the damage of the input assignment. Target Penalty function ensures that all target compounds are inhibited [see Constraint (3)]. Reaction Inference Penalty function controls the inhibition condition of reactions [see Constraint (4)]. Compound Inference Penalty function ensures the inhibition condition of compounds [see Constraint (5)]. Next, we elaborate on construction of these functions, and explain how they fit to corresponding constraints.

2.2.1 Damage scoring function

This function models the toxicity arising from the disturbance of the production of those compounds that are not intended to be inhibited, but are inhibited as a result of inhibiting a subset of enzymes in the given network G. We compute the damage scoring function in terms of the binary variables x_c ($c \in C$) with a positive constant k_1 as (see Lemma 1 in Supplementary Material S1—SM 4):

$$H_{\text{damage}} = k_1 \sum_{c \in C - C_{\text{target}}} x_c \tag{6}$$

2.2.2 Target penalty function

This function models the loss in the efficacy of the drug by representing the constraint that requires inhibition of all the compounds in the target set [see Constraint (3)]. We write this function in terms of the binary variables x_c above with a positive constant k_2 as follows:

$$H_{\text{target}} = k_2 \sum_{c \in C_{\text{target}}} (1 - x_c) \tag{7}$$

The target penalty H_{target} in (7) is minimized if and only if the states of all compounds in the target set are 1 (i.e. when all the targeted compounds are inhibited). In other words, the function H_{target} only returns minimum value for valid assignment of $\mathbf{x}$ (see Lemma 2 in Supplementary Material S1—SM 4).

2.2.3 Reaction inference penalty

This function expresses the second constraint in the inhibition conditions [see (4)]. Given a reaction $r \in R$, we construct a system of two linear inequalities from (4) as:

$$x_r \geq x_v \forall v \in N(r) \tag{8}$$

$$0 \geq x_r - \sum_{v \in N(r)} x_v \tag{9}$$

Since each variable x_r above takes value either 0 or 1, in order to satisfy inequality (8), $\forall v \in N(r)$, we must have $x_r - x_v = 0$, or $x_r - x_v - 1 = 0$. Following from these two observations, we construct a quadratic expression for inequality (8) as follows:

$$\sum_{v \in N(r)} [(x_r - x_v)^2 + (x_r - x_v - 1)^2] - |N(r)| \tag{10}$$

Expression (10) obtains the minimum value of 0 if and only if x_u, and x_v with $v \in N(u)$ satisfy inequality (8).

Building a quadratic equation for modeling inequality (9) is nontrivial. This is because the value of difference $(x_r - \sum_{v \in N(r)} x_v)$ depends on the number of nodes in $N(r)$. To overcome this challenge, we define $|N(r)| + 1$ auxiliary binary variables $t_{r0}, t_{r1}, \ldots$, and write the following expression:

$$\left[x_r - \sum_{v \in N(r)} x_v + \sum_{\alpha=0}^{|N(r)|} (\alpha t_{r\alpha})\right]^2 + (1 - \sum_{\alpha=0}^{|N(r)|} t_{r\alpha})^2 \tag{11}$$

In Expression (11), each of the $|N(r)| + 1$ auxiliary binary variables $t_{r\alpha}$ models one of the $|N(r)| + 1$ possible valid values the right-hand side of inequality (9) can take. Therefore, Expression (11) reaches to value of 0 if and only if states x_r and x_v, with $v \in N(r)$ satisfy inequality (9), and the variable $t_{r\alpha} = 1$ only if $x_r - \sum_{v \in N(r)} x_v + \alpha = 0$.

Using Expressions (10) and (11) for reaction $r \in R$, we construct the reaction inference penalty function which ensures the second constraint from the inhibition conditions using a tunable positive constant k_3, and constant λ_R as:

$$\begin{aligned} H_{\text{reaction}} = k_3 \sum_{r \in R} \Bigg\{ & \sum_{v \in N(r)} [(x_r - x_v)^2 + (x_r - x_v - 1)^2] \\ & + \left[x_r - \sum_{v \in N(r)} x_v + \sum_{\alpha=0}^{|N(r)|} (\alpha t_{r\alpha})\right]^2 + (1 - \sum_{\alpha=0}^{|N(r)|} t_{r\alpha})^2 \Bigg\} + \lambda_R \end{aligned} \tag{12}$$

The function H_{reaction} returns minimum value of 0 only for valid assignment of $\mathbf{x}$, and auxiliary variable t (see Lemma 3 in Supplementary Material S1—SM 4).

2.2.4 Compound inference penalty

This function models the third, and the final constraint in the inhibition conditions [see (5)]. Given a compound $c \in C$, we construct a system of two linear inequalities from (5) as:

$$x_c \leq x_v \forall v \in N(c) \tag{13}$$

$$-|N(c)| \leq x_c - \sum_{v \in N(c)} x_v - 1 \tag{14}$$

In order to satisfy inequality (13), $\forall v \in N(c)$, we must have $x_c - x_v = 0$, or $x_c - x_v + 1 = 0$. These two observations lead to the following quadratic expression:

$$\sum_{v \in N(c)} [(x_c - x_v)^2 + (x_c - x_v + 1)^2] - |N(c)| \tag{15}$$

Expression (15) yields minimum value of 0 if and only if x_c and $\forall v \in N(c)$, x_v satisfy inequality (13). The proof of this statement is the same as that for (10).

To build a quadratic expression for inequality (14), similar to Expression (11), we define $|N(c)| + 1$ auxiliary binary variables $w_{c0}, w_{c1}, \ldots$. We have the quadratic expression for inequality (14) as follows:

$$\left[x_c - \sum_{v \in N(c)} x_v - 1 + \sum_{\beta=0}^{|N(c)|} (\beta w_{c\beta})\right]^2 + (1 - \sum_{\beta=0}^{|N(c)|} w_{c\beta})^2 \tag{16}$$

We observe from Expression (16) that $|N(c)| + 1$ auxiliary binary variables $w_{c\beta}$ correspond to $|N(c)| + 1$ possible valid values the right-hand side of inequality (14) can take. Therefore, Expression (16) takes minimum value of 0 if and only if states x_c and x_v satisfy inequality (14), and the variable $w_{c\beta} = 1$ only if $x_c - \sum_{v \in N(c)} x_v - 1 + \beta = 0$.

Using Expressions (15) and (16) for compound $c \in C$, we construct the compound inference penalty with the help of a tunable positive constant k_4, and constant λ_C as:

$$\begin{aligned} H_{\text{compound}} = k_4 \textstyle\sum_{c \in C} \Bigg\{ & \sum_{v \in N(c)} [(x_c - x_v)^2 + (x_c - x_v + 1)^2] \\ & + \left[x_c - \sum_{v \in N(c)} x_v - 1 + \sum_{\beta=0}^{|N(c)|} (\beta w_{c\beta})\right]^2 + (1 - \sum_{\beta=0}^{|N(c)|} w_{c\beta})^2 \Bigg\} + \lambda_C \end{aligned} \tag{17}$$

The function H_{compound} returns minimum value of 0 only for valid assignment of $\mathbf{x}$, and auxiliary variable w (see Lemma 4 in Supplementary Material S1—SM 4).

Combining functions (6), (7), (12), and (17), we represent the TIE problem in the QUBO form by an energy function as follows:

$$H = H_{\text{damage}} + H_{\text{target}} + H_{\text{reaction}} + H_{\text{compound}} \quad (18)$$

We observe that the function H is a combination of linear and quadratic forms. Because binary variable x satisfies $x = x^2$, we rewrite the linear term x as x^2 making the entire of equation quadratic. As a result, the function H is in the QUBO form, which can be processed by QA. The final result that we expect to obtain after QA process is a set of variables $Y^\star = X \cup \{t_{r\alpha} | r \in R, \alpha \in N(r)\} \cup \{w_{c\beta} | c \in C, \beta \in N(c)\}$ such that:

$$Y^\star = \text{argmin}\{H\} \quad (19)$$

The function H models the TIE problem because it satisfies the two principles we mentioned before in this section. Penalty functions H_{target}, H_{reaction}, and H_{compound} return the minimum value of 0 for only valid assignments. As a result, if we choose positive constants k_2, k_3, k_4 that are large enough, the outputs of the function H for valid assignments are always lower than those for invalid assignments. In addition, the function H_{damage} returns the damage corresponding to the input assignment if we set $k_1 = 1$. Thus, for valid assignments whose penalty scores are always equal to 0, the function H returns corresponding damage of those valid assignments (see Theorem 1 in Supplementary Material S1—SM 4).

3 Discussion

In this section, we evaluate our method on a small dataset and a large dataset. We describe datasets in detail in Supplementary Material S1—SM 3. We compare our method to four methods:

- **Exact method (OPMET):** This method uses branch-and-bound to examine all possible combinations of inhibited enzymes, and thus it is optimal (Sridhar *et al.* 2008).
- **Integer programming (IP):** We use the IP formulation for the *BN-ReactionCut* problem (Tamura *et al.* 2010) with a modification in its objective function to solve TIE problem. *BN-ReactionCut* problem makes a simplifying, but incorrect assumption that each reaction is controlled by one enzyme. As a result, it ignores the set of enzymes operating on the metabolic network. In order to make fair comparisons with our method, we post-process the solution returned by IP, and randomly select one enzyme corresponding to each inhibited reaction from the resulting solution for inhibition. We then calculate actual damage caused by inhibiting selected enzymes. For each test case, we perform process for a constant number of times, and report the average damage.
- **Heuristic (double iterative):** This is a heuristic variant of OPMET working in two phases (Song *et al.* 2009).
- **Simulated annealing (SA):** This method is inspired by the annealing process, similar to QA. However, unlike QA, SA works on a classical computer. We run SA with the same objective function H that we use in QA.

3.1 Evaluation using synthetically selected targets

Our first set of experiments answer the question: How does QuTIE perform under different network characteristics and number of target compounds compared to existing solutions?

3.1.1 Experimental settings

3.1.1.1 Experimental setup for Quantum Hybrid Solver

We use a Quantum Hybrid Solver provided by D-Wave to solve our proposed QUBO. It is a hybrid framework combining classical and quantum computing techniques to find optimal solutions for a given QUBO formulation. We explain this framework in detail in Supplementary Material S1—SM 2. We set the running time limit of the Quantum Hybrid Solver to 10 min for the small datasets, and 20 min for the large ones. In all experiments, solutions provided by Quantum Hybrid Solver are valid (i.e. all target compounds are eliminated).

3.1.1.2 Experimental setup for target selection

Let us denote the number of target compounds to be inhibited with k. Given a metabolic network, we run experiments by growing the number of target compounds to be inhibited in that network from $k = 2$ to 27, at increments of 5 (i.e. six different values of k) by randomly selecting k target compounds from that network. We repeat this procedure up to five times for each combination of metabolic network and target network size, measure damage and running time, and report the average.

3.1.1.3 Experimental setup for datasets

We use metabolic pathways for three species: *Escherichia coli* (eco or *E.coli*), *Homo sapiens* (hsa or *H.sapiens*), and *Mus musculus* (mmu or *M.musculus*) from the KEGG database (Kanehisa and Goto 2000). We categorize these metabolic networks into two groups based on the number of interactions in each: small and large pathways. The number of nodes in small pathways ranges from 35 to 93, while the number of nodes in large pathways ranges from 146 to 305. Supplementary Tables S1 and S2 list the characteristics of pathways in more detail.

3.1.2 Comparison with the exact method

Here, we examine the performance of QuTIE on small datasets by comparing to the exact method, OPMET. In the small datasets, we do not include the *Pyruvate* metabolic network of *E.coli*, and the *Glycine, serine, and threonine* metabolic networks of *H.sapiens* and *M.musculus* because their sizes are too big for the exact method to run. Notice that the exact method guarantees optimal solutions. Thus, the fundamental purpose of this comparison is to observe (i) how well QuTIE optimizes the damage function for the TIE problem under different metabolic networks in the small datasets, and various target compound set sizes, with respect to the optimal solution, and (ii) how much running time QuTIE needs to arrive at this solution as compared to the exact method. We use the small networks listed in Supplementary Table S1 for the exact method does not scale to larger networks. In total, we perform 900 experiments (i.e. 5 networks $\times$ 3 species $\times$ 6 values of $k \times$ 5 random repetitions $\times$ 2 methods).

Recall that the TIE problem aims to minimize the damage, while inhibiting all target compounds. We first compare the two methods in terms of the damage their results inflict on the

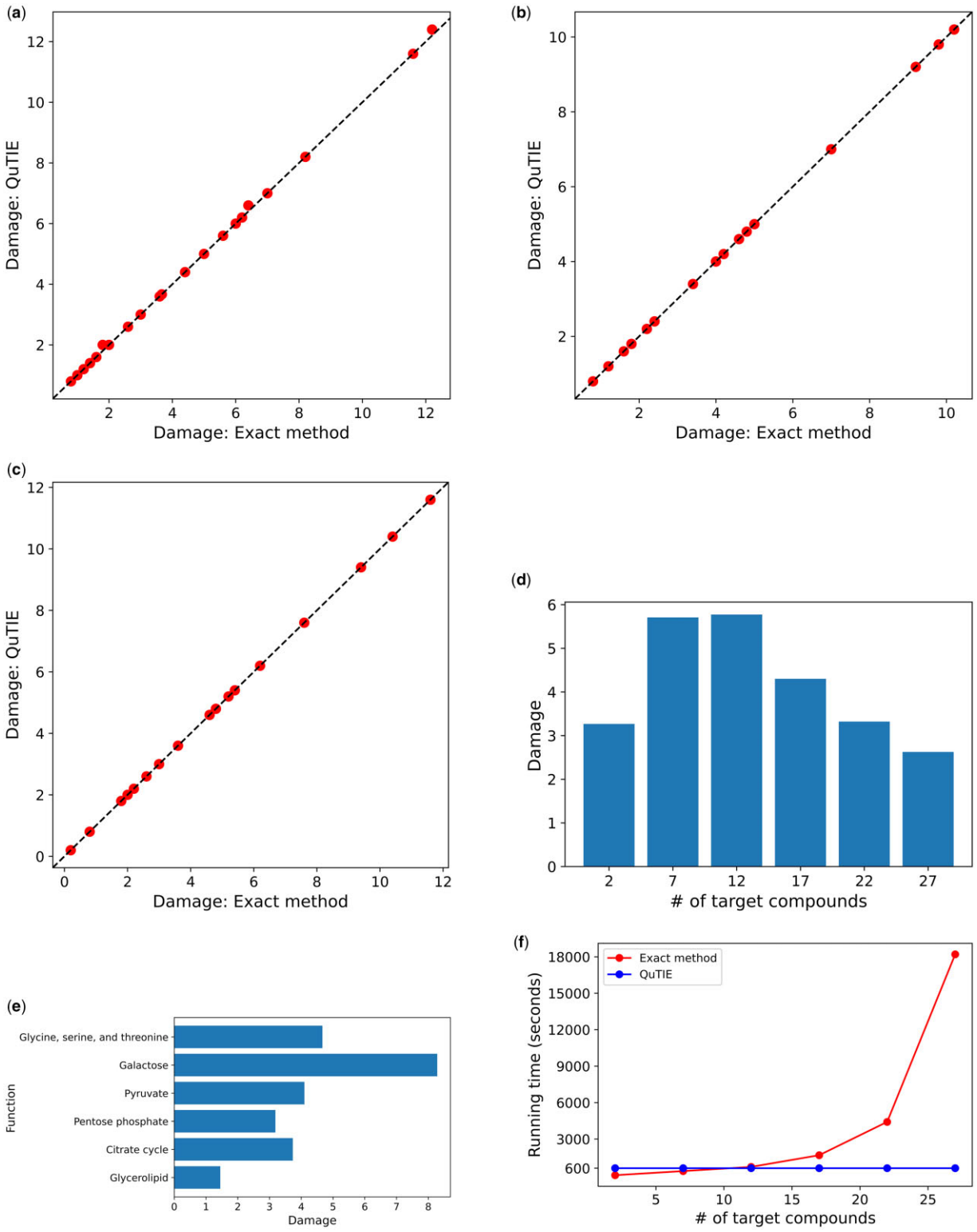

Figure 1. Analysis of QuTIE on small datasets. (a–c) Damage values provided by QuTIE and the exact method for the three species *E.coli*, *H.sapiens*, and *M.musculus*, respectively. Each point corresponds to the average of a combination of one network and one k value across all test cases. The diagonal line is the $x = y$ line. (d) Average damage value of QuTIE across all parameters grouped by the number of target compounds. (e) Average damage value of QuTIE across all parameters grouped by the network function. (f) Comparison between QuTIE and the exact method on small datasets in terms of running time.

given metabolic network. We obtain the average damage of solutions from QuTIE and the exact method for each combination of metabolic networks and target compound set size. Figure 1a–c presents the results for three species including *E.coli*, *H.sapiens*, and *M.musculus* respectively. Each point in this figure corresponds to the average damage of one (network, target compound set size) pair. Our results demonstrate that QuTIE is able to obtain the optimal or near optimal

solutions for all experimental settings. Figure 1 illustrates the average damage of experimental settings for the same network function. From the results, we observe that inhibiting target compounds from galactose metabolic networks can cause more damage than those from any other functions. In addition, we examine the average damage for different number of target compounds, and show the results in Fig. 1. We observe an upward trend in the average damage when the number of target compounds is small. We infer that inhibiting medium-sized sets of target compounds may cause the most damage to the network.

One of the fundamental promises of quantum optimization algorithms is that they can solve problems with high complexity dramatically faster than the algorithms operating on traditional computers. Following from this, and our observation above that QuTIE yields optimal damage values even for very large values of k, the next important question we need to answer is at what running time cost does our algorithm achieve these results, for the TIE problem is NP-hard?

We compare the running time of the two methods. The running time of QuTIE is the time limit set in the Quantum Hybrid Solver (10 min) in all experiments for the small dataset. For the exact solution, we report the average running time for each size of target compound set. We report the running time of two methods in Fig. 1. Our results suggest that the number of target compounds k has massive impact on the total running time of the exact method on classical computers. This is expected as the complexity of the TIE problem is exponential in the target compound set size in the worst case. For small k, finding optimal solutions is trivial. As the value of k increases though, the running time to find exact solutions quickly becomes impractical. The second observation is that QuTIE, on the other hand, is not affected by the value of k, and it yields optimal solutions in the preset time limit. Although we cannot claim anything about precise quantum speed-up over classical exact method due to fixed time limit, quantum computing shows its potential power for reaching optimal solutions in a fast manner. To be clear, in the case with $k = 27$, QuTIE can find optimal solutions in a duration that is only equal to nearly 3% of the running time of the exact method. The gap between the running time of our method and that of the exact solution grows with increasing value of k.

3.1.3 Comparison with integer programming solution

Here, we examine the performance of QuTIE on small datasets by comparing it against the IP method that is one of the most popular method for solving optimization problems like TIE (Tamura et al. 2010).

We compare QuTIE and IP in terms of damage from resulting solutions. Figure 2a–c presents the results for three species including *E.coli*, *H.sapiens*, and *M.musculus*, respectively. Each point in this figure corresponds to the average damage of one (network, target compound set size) pair. The results demonstrate that QuTIE can provide solutions with less damage than IP in most cases (99.3% of all test cases). Recall that the IP formulation does not consider the relation of enzymes and reactions at first glance, while in QuTIE we present the dependency of enzymes and reactions in (12). Thus, the results imply that taking the set of enzymes into account is crucial in optimizing damage for the TIE problem.

3.1.4 Comparison with the double iterative method

Next, we study the performance of QuTIE on larger datasets. Exact method does not scale to these networks, and so we compare our method to the heuristic double iterative method. Our goal in this experiment is to observe whether the damage incurred by the solutions of our method are better than existing heuristic solutions, which also scale to large networks and large values of k. We use the networks listed in Supplementary Table S2 (see Supplementary Material S1—SM 3). In total, we perform 288 experiments (i.e. 4 networks $\times$ 3 species $\times$ 6 values of $k\times$ 2 random repetitions $\times$ 2 methods).

Figure 3 plots the damage values resulting from the two methods for each combination of network and target compound set size. We observe that QuTIE method outperforms the heuristic solution in almost all cases; QuTIE identifies a solution with less than or equal damage than that found by the heuristic method in 117 out of 144 cases (i.e. in 81.2% of the experiments). In 12.6% of the experiments, the gap between our method and the heuristic solution is more than 4 in favor of our method, while the heuristic method never yields damage gap > 4 in any of the cases. These results suggest that QuTIE has the potential to identify target enzymes for even large networks when the exact methods do not work without relying on heuristics. The points outside the zone bounded between two lines $y = x + 4$, and $y = x - 4$ are from nucleotide metabolism networks, and purine networks. This suggests that the underlying network topology has great influence on how much QuTIE outperforms the competing method.

3.1.5 Comparison with simulated annealing

Here, we compare the performance of QuTIE to its counterpart, which is executed on a classical computer on small

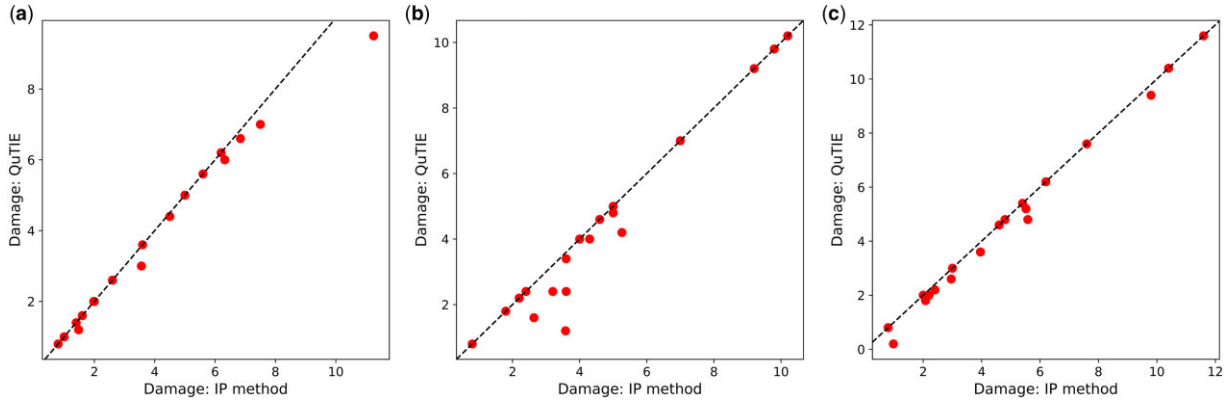

Figure 2. Analysis of the IP method and QuTIE. (a–c) Damage values for the three species *E.coli*, *H.sapiens*, and *M.musculus*, respectively. Each point corresponds to the average of a combination of one network and one k value across all test cases. The diagonal line is the $x = y$ line.

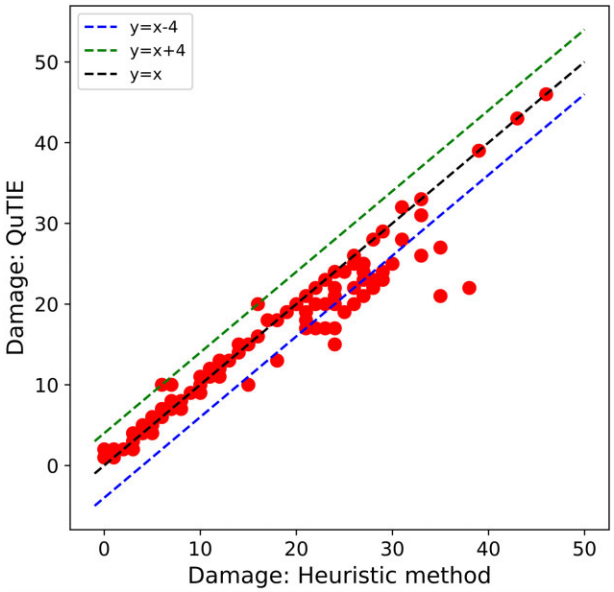

Figure 3. Comparison between QuTIE and the heuristic double iterative method in large datasets in terms of damage. The less damage is, the better solution is. Data points outside the envelope formed by the green dash line, and the blue dash line indicate cases in which QuTIE significantly outperforms the heuristic method.

datasets. Our goal in this experiment is to observe that given a same objective function, whether a quantum computer can explore better solutions through annealing process than would a classical computer. Recall that a solution is valid if it inhibits all target compounds. We compare QuTIE and SA in terms of the number of times a valid solution can be found over the total number of test cases. We set the time limit for both methods to 10 min. Figure 4 shows the percentage of valid solutions for the methods for different datasets and increasing number of target compounds k. We observe that QuTIE always finds valid solutions. Meanwhile, SA rarely can find valid solutions in cases of $k > 2$ (<20%). Even in the cases of small target size ($k = 2$), SA fails to find valid solutions in 50–60% of test cases. The results imply that quantum computers can outperform classical computers in solving the TIE problem by simulating annealing process.

3.2 The impact on actual disease-related compounds

So far, we tested our method on real networks, but with randomly selected target compound sets. Here, we evaluate how our method performs when the target compound sets are verified to be associated with known disease classes. We use the biosynthesis of amino acids metabolic network of *H.sapiens* for every experiment in this part. We obtain mapping from disease classes to compounds from the literature (Zielinski *et al.* 2015), for 14 major disease classes. Figure 5 shows these disease classes and the number of compounds associated for each disease class in the given metabolic network. We observe that the number of target compounds shows huge variation among different disease classes (it varies from 1 to 10 with a median of 4). This illustrates the need for new solutions that work well for both small and large target sets.

We run QuTIE for each disease class using its corresponding associated compounds as the target compound set. We report the damage as well as the target enzymes identified by our method. Figure 5 shows the damage QuTIE yields for

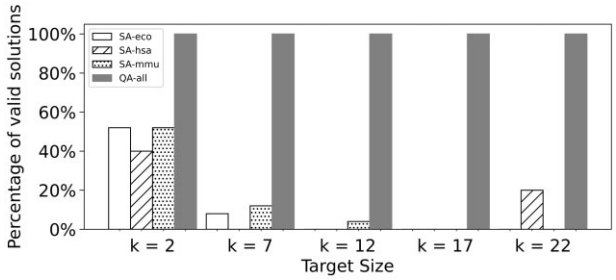

Figure 4. Comparison between QuTIE and the SA method on small datasets in terms of their success in finding valid solutions.

each disease class. Similar to Fig. 5, we observe a huge variation in the damage value (from 2 to 17, with a median of 10). The values in these two figures, however, are not correlated. That is, smaller target compound set size does not necessarily yield smaller damage. This suggests that the topology of the metabolic network and the distribution of the target compounds over this topology play an important role in the efficacy and the toxicity of the drugs designed for the underlying disease. For example, Fig. 5a and b together suggests that disease classes bleeding, digestive, and immune deficiencies can be treated with less damage, although they have more target compounds than several other disease classes, such as pain, urinary, and liver failures.

We also examine the number of inhibited enzymes by QuTIE for each disease class. This number can be considered as an indicator of the cost/difficulty of inhibiting the enzymes needed to stop the production of those compounds: The larger the number of enzymes, the more effort it takes to inhibit them. Figure 5 presents the results. Similar to Fig. 5a and b, we observe a huge variation in the number of enzymes (ranging from 14 to over 40). It is worth noting that QuTIE aims to minimize the damage, and not the number of enzymes. Therefore, it is not surprising that QuTIE can sometimes yield a very large number enzymes to obtain smaller damage. We also observe no correlation between the target compound set size and the number of enzymes resulting from them. This also implies that the topology of the distribution of target compounds on the metabolic network is the primary factor in the size of the resulting enzyme set.

3.2.1 Evaluation of disease enzyme associations

Our final analysis explores whether the targeted enzymes indeed have known associations for the disease classes, whose compounds they inhibit. To do that, we list all the enzymes that QuTIE identifies as target in order to inhibit the target compounds associated for each disease class. For 14 disease classes, and all the known compounds associated with each disease class, QuTIE identifies 408 enzymes as targets in total with repetition (i.e. an enzyme can be a target for multiple disease classes), leading to 53 unique enzymes for all disease classes combined. We provide the list of disease classes, compounds, and enzymes in Supplementary Material S2. Further studying these enzymes reveals that QuTIE indeed identifies target genes verified to be affecting the target disease in wet-lab experiments on human as well as different animal models. Table 1 lists five examples out of these due to page limitations.

The very first target enzyme that we identify for the pain category is catechol O-methyltransferase. Studies on both rat and mice models demonstrate that the inhibition of catechol O-methyltransferase affects the perception of pain (Kambur

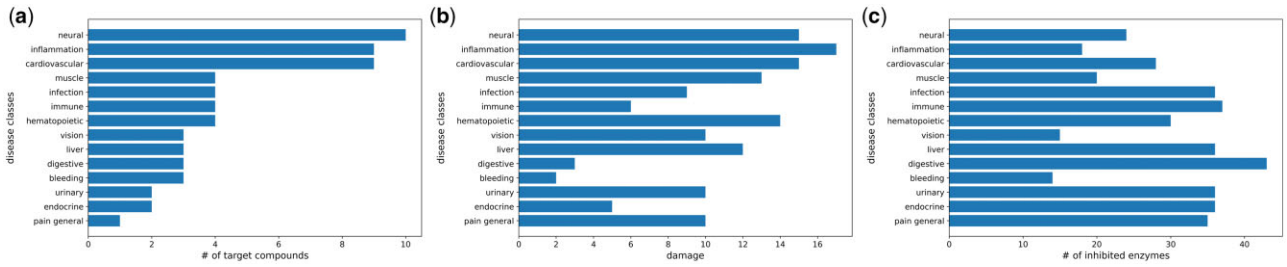

Figure 5. Analysis of the QuTIE in cases of disease-related target compounds in the biosynthesis of amino acids metabolic network. (a) Summary of the number of target compounds related to different disease classes. (b) The correlation between disease classes, and average resulting damage. (c) The correlation between disease classes, and the average number of inhibited enzymes.

Table 1. Five example target enzymes identified by QuTIE for five disease classes and publication evidences for those enzymes.

Disease class	Enzyme name	Evidence
Pain general	Catechol O-methyltransferase	Kambur and Mannisto (2010)
Endocrine	Aspartate aminotransferase	Barse *et al.* (2007)
Urinary	Arginase	Naylor and Cederbaum (1981)
Bleeding	Prolin oxidase	Pandhare *et al.* (2009)
Immune	Glutamine synthetase	Cruzat *et al.* (2018)

and Mannisto 2010). Similarly, the first target enzyme that we identify for the endocrine disease category is aspartate aminotransferase, whose altered activity in children has a detrimental effect of early-life endocrine-disrupting chemical exposure on liver function (Barse *et al.* 2007). QuTIE identified arginase as one of the top target enzymes for urinary disorders. Indeed, in clinical samples of obstructive nephropathy, altered levels of arginase was observed using both Western blot and MRM analyses (Naylor and Cederbaum 1981). Hyperprolinemia is a bleeding disorder, caused by the build-up of proline in the blood (Pandhare *et al.* 2009). QuTIE correctly identifies this enzyme too as a potential drug target. Finally, the abundance of glutamine, one of the targets we identify for immune-related disorders, is indeed linked to the immune suppression in humans (Cruzat *et al.* 2018). In summary, there is substantial publication evidence supporting the potential target enzymes that we identify, which suggest that efficient and accurate solution to the TIE problem on large and complex networks using quantum optimization has great potential to assist drug target identification.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work is partially supported by the National Science Foundation under grants number 2111679 and 1908594.

Data availability

The datasets were derived from the following public domain resources: https://www.genome.jp/kegg/pathway.html and in the Supplementary Materials.

References

Barse AV, Chakrabarti T, Ghosh TK *et al.* Endocrine disruption and metabolic changes following exposure of *Cyprinus carpio* to diethyl phthalate. *Pestic Biochem Physiol* 2007;**88**:36–42.

Berillo D, Yeskendir A, Zharkinbekov Z *et al.* Peptide-based drug delivery systems. *Medicina* 2021;**57**:1209.

Carman GM, Han G-S. Phosphatidic acid phosphatase, a key enzyme in the regulation of lipid synthesis. *J Biol Chem* 2009;**284**:2593–7.

Choi V. Minor-embedding in adiabatic quantum computation: I. The parameter setting problem. *Quantum Inf Process* 2008;**7**:193–209.

Cohen J, Powderly W, Opal S. *Infectious Diseases*. 3rd edn. Amsterdam, Netherlands: Elsevier Inc., 2010.

Cooney M. Kinetic measurements for enzyme immobilization. *Methods Mol Biol* 2017;**1504**:215–32.

Copeland RA, Harpel MR, Tummino PJ. Targeting enzyme inhibitors in drug discovery. *Expert Opin Ther Targets* 2007;**11**:967–78.

Cruzat V, Macedo Rogero M, Noel Keane K *et al.* Glutamine: metabolism and immune function, supplementation and clinical translation. *Nutrients* 2018;**10**:11.

D'Avanzo F, Rigon L, Zanetti A *et al.* Mucopolysaccharidosis type II: one hundred years of research, diagnosis, and treatment. *IJMS* 2020; **21**:1258.

Fox DM, MacDermaid CM, Schreij AMA *et al.* RNA folding using quantum computers. *PLoS Comput Biol* 2022;**18**:e1010032.

Guillaume A, Goh EY, Johnston MD *et al.* Deep space network scheduling using quantum annealing. *IEEE Trans Quantum Eng* 2022;**3**:1–13.

Jerbi S, Gyurik C, Marshall SC *et al.* Parametrized quantum policies for reinforcement learning. In: Beygelzimer A, Dauphin Y, Liang P, Wortman Vaughan J (eds), *Advances in Neural Information Processing Systems*, Vol. 34. Red Hook, NY, USA: Curran Associates, Inc., 2021, 28362–75.

Kambur O, Mannisto PT. Catechol-o-methyltransferase and pain. In: Nissinen E (ed.), *International Review of Neurobiology*, Vol. 95. Cambridge, MA: Academic Press, 2010, 227–79.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.

Li M, Zhao G, Su W-K *et al.* Enzyme-responsive nanoparticles for anti-tumor drug delivery. *Front Chem* 2020;**8**:647.

Martorana A, Koch G. Is dopamine involved in Alzheimer's disease? *Front Aging Neurosci* 2014;**6**:252.

Mizutani S, Pauwels E, Stoven V *et al.* Relating drug–protein interaction network with drug side effects. *Bioinformatics* 2012;**28**:i522–8.

Mulligan VK, Melo H, Merritt HI *et al.* Designing peptides on a quantum computer. bioRxiv, 2019. https://doi.org/10.1101/752485. preprint: not peer reviewed.

Nałęcz-Charkiewicz K, Nowak RM. Algorithm for DNA sequence assembly by quantum annealing. *BMC Bioinformatics* 2022;**23**:122.

Naylor EW, Cederbaum SD. Urinary pyrimidine excretion in arginase deficiency. *J Inherit Metab Dis* 1981;**4**:207–10.

Pandhare J, Donald SP, Cooper SK *et al.* Regulation and function of proline oxidase under nutrient stress. *J Cell Biochem* 2009;**107**:759–68.

Preskill J. Quantum computing in the NISQ era and beyond. *Quantum* 2018;**2**:79.

Pritchard AL, Ratcliffe L, Sorour E *et al.* Investigation of dopamine receptors in susceptibility to behavioural and psychological symptoms in Alzheimer's disease. *Int J Geriatr Psychiatry* 2009;**24**:1020–5.

Riley AL, Kohut S. *Drug Toxicity*. Berlin/Heidelberg: Springer Berlin Heidelberg, 2010.

Robertson JG. Mechanistic basis of enzyme-targeted drugs. *Biochemistry* 2005;**44**:5561–71.

Robertson JG. Enzymes as a special class of therapeutic target: clinical drugs and modes of action. *Curr Opin Struct Biol* 2007;**17**:674–9.

Shankarappa SA, Koyakutty M, Nair SV *et al.* Efficacy versus toxicity - the ying and yang in translating nanomedicines. *Nanomater Nanotechnol* 2014;**4**:23.

Shor PW. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Rev* 1999;**41**: 303–32.

Song B, Sridhar P, Kahveci T *et al.* Double iterative optimisation for metabolic network-based drug target identification. *Int J Data Min Bioinform* 2009;**3**:124–44.

Song B, Büyüktahtakın IE, Bandyopadhyay N *et al.* Identifying enzyme knockout strategies on multiple enzyme associations. In: Mahdavi MA (ed.), *Bioinformatics*, Chapter 17. Rijeka: IntechOpen, 2011.

Sridhar P, Song B, Kahveci T *et al.* Mining metabolic network for optimal drug targets. In: *Pacific Symposium on Biocomputing*. Hackensack, NJ, USA: World Scientific, 2008, 291–302.

Tamura T, Takemoto K, Akutsu T. Finding minimum reaction cuts of metabolic networks under a Boolean model using integer programming and feedback vertex sets. *Int J Knowl Discov Bioinform* 2010; **1**:14–31.

Terentis AC, Freewan M, Sempértegui Plaza TS *et al.* The selenazal drug ebselen potently inhibits indoleamine 2,3-dioxygenase by targeting enzyme cysteine residues. *Biochemistry* 2010;**49**:591–600.

Zielinski DC, Filipp FV, Bordbar A *et al.* Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis. *Nat Commun* 2015;**6**:7101.