Exploration of bioinformatic domain based on data mining, reaction and enzyme promiscuity predictions

Chonghuan Zhang,[1] *Qianyue Zhang*[1] *and Alexei A. Lapkin*[1, 2 *]

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom

[2] Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd, 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore

∗ Corresponding author E-mail addresses: aal35@cam.ac.uk

Abstract

Biochemical transformations may allow significant improvements in synthetic efficiency of complex functional molecules through reduction in the number of synthetic steps or avoidance of harsh conditions and/or toxic solvents/reactants. Yet, there is a limited access to biochemical reaction data, which reduces the opportunities of finding alternatives and discovering synergies with organic synthesis. We propose a workflow to explore the sparse synthetic biological domain. Using a molecular graph method we predict feasible biosynthetic reactions. The products of biosyntheses are learned from the functional transformations of the literature-excerpted reactions recorded in KEGG database. Through this approach we expanded the KEGG reaction dataset of biochemical transformations by approximately four times. To catalyse the novel reactions, we proposed a transformer model that learns from reaction SMILES and amino acid sequences of native enzymes and predicts promiscuous enzymes for potential substrates. The proposed transformer model calibrates the feasibility of the predicted reactions and reduces the search scope for promiscuous enzymes in the pool. A populated biological reaction space is eventually visualised in a two-dimensional t-SNE diagram.

***Keywords**:* bioinformatics; reaction network; synthetic biology, data mining, machine learning

Introduction

Conventional chemical products development starts from raw materials, such as petrochemical feedstocks or bio-based resources. It involves chemical synthesis steps to obtain the desired target intermediates (industrial chemicals) and functional molecules.[1, 2] In contrast, synthetic biology is concerned with molecules that are intermediates at various stages of metabolic pathways, and could lead to production of target substances in cell-based bioreactors. Compared to organic synthetic paths, the bio-synthetic paths typically have higher redox efficiency and

moderate reaction conditions; both factors are desirable from the point of view of reducing environmental impact of industrial processes (we do not discuss here the well-known issues of complexity of product separation from bio-processes).[3, 4] An addition of synthetic biology into computer-assisted synthesis planning (CASP) of pharmaceuticals and industrial chemicals certainly opens up opportunities for more efficient chemical production.[5-7]

Recently we generated a hybrid reaction network to bridge the domains of organic synthesis and synthetic biology.[5] From the hybrid network, it is clear that the biological reaction space is very sparse compared with the organic chemical network, involving only 0.35% of the total reactions in the combined dataset. Similarly, Levin *et al.*[8] published a method of merging enzymatic and synthetic chemistry to guide CASP. In their work, reaction templates were summarised from the reaction transformations, and hybrid synthetic pathways were assembled from reaction templates. Among the 171 thousand reaction templates, only 4.6% were from enzymatic templates. Both practical implementations of CASP based on both bio-catalytic and chemical synthesis were hindered by the limited amounts of the biochemical reactions data.

To enrich the biological reaction dataset, several methods were proposed to predict possible biological reactions. Some research groups extracted reaction mechanisms and reproduced fragmentations of the reactions, such as ReactPRED,[9] BioTransformer,[10] and SyGMa.[11] Others used mechanism-free machine learning methods to learn from past reactions, such as GLORYx[12] and Litsa *et al.*'s method.[13, 14] Jiang *et al.*[15] learned from enzymatic reaction network connectivity, and predicted enzymatic reactions using link prediction. These methods are likely to generate metabolites and metabolic reactions which are unseen in prior literature, but with low confidence; besides, few of them provide information about the enzymes required. Therefore, we propose to learn from the literature-excerpted biological reactions and predict novel reactions to expand the biological reaction dataset. The feasibility of the predicted

reactions was calibrated with a natual language processing (NLP) transformer to suggest promiscuous biocatalysts for the reactions. We hypothesise that by diversifying the biological reaction dataset, the potential for synthetic biology to produce molecules of interests would be increased.

Novel reaction pathways can be developed/discovered with tailored enzymes, the biocatalysts. Enzymes have been developed through rational design or directed evolution. Rational design is a top-down method which considers enzyme behaviour and functionality, and makes proteins from scratch by predicting their gene sequences that fold to specific structures;[16] directed evolution is a bottom-up method that screens from multiple experimentally evolved protein structures to satisfy specific objectives.[17] These two methods are usually combined to accelerate enzyme development.[18]

Apart from conventional approaches, machine learning is also frequently applied to enzyme engineering. Most recently, Google's DeepMind published its open-sourced deep learning protein structure prediction tool AlphaFold,[19] and then its upgraded version AlphaFold2,[20] which use new neural network architectures to learn from three-dimensional structures of proteins and make predictions at a near-experiential precision. A comprehensive discussion of machine learning applications in enzyme engineering can be found in a review paper by Mazurenko *et al*.[21]

Apart from evolving novel enzymes, synthetic biologists also benefit from enzyme promiscuity, which is the capability of enzyme's protein structure to bind non-native substrates and catalyse multiple reactions.[22, 23] Enzyme promiscuity is efficient in exploring enzyme's substrate

specificity, since at least one third of protein superfamilies are functionally diverse and each superfamily is able to catalyse more than one reaction.[24]

The development of a promiscuous enzyme suggestion model relies on NLP deep learning. NLP methods have experienced a period of rapid development over the last few years. Notable methods are long short-term memory (LSTM) neural network,[25] transformer,[26] bidirectional encoder representations from transformers (BERT),[27] *etc*. We should note that NLP is particularly promising in bioinformatics due to the natural fit to convert enzymes into amino acid letters and reactions into machine readable string representations. For reaction prediction, Kreutter *et al*.[28] developed an enzymatic transformer, to convert enzymatic reactions into reaction SMILES and enzymes into language tokens. The transformer predicts reaction products with remarkable accuracy. For biosynthetic planning, Probst *et al*.[29] generalised Molecular Transformer,[30] a deep learning reaction prediction transformer model, to predict biocatalytic reaction outcomes and build pathways. Ofer *et al*.[31] discussed the possibility to use the protein sequence, the amino acid chain, as a language and reviewed numbers of protein-related tasks solved by NLP methods. NLP certainly can provide insights to tackle bioinformatics problems.

In this work we propose a workflow to expand the knowledge of biochemical reactions, specifically based on data mining from KEGG database, and to build a new layer of information on our previous work[5] of hybrid chemical and biological reaction network. This starts from prediction of novel metabolic reactions from a molecular graph-based method to learn from functional transformations of analogue reactions in the reaction dataset and to populate the original biological reaction network. To catalyse the novel predicted reactions, instead of designing new enzymes, which are case-specific, we use an NLP transformer to learn from

enzyme language, amino acid sequences, with no segmentation on amino acid chain, and reaction language, reaction SMILES strings, and suggest promiscuous enzymes to bind with the substrates in the predicted reactions, and the suggestions are given with binding possibilities. The populated biological space would be analysed to investigate its feasibility of conducting the proposed bio-chemical transformations.

Methods

Synthetic biological reaction domain data mining

All reactions from KEGG[32] reaction database were mined. Since enzymes catalyse metabolic reactions from both directions, all reactions were assumed to be reversible.[15] Therefore, for all metabolic reactions, both reaction directions were considered in the local reaction database. The reactions record all reaction participants, including substrates, products, cofactors such as ATP and NADPH, and free metabolites such as oxygen.

In retrosynthesis planning, a target molecule is broken down by chains of reactions, until all intermediates reach their precursors. Cofactors and free metabolites in most cases always exist as reaction intermediates, and do not usually contribute to the carbon flow in reaction. Therefore, in the local reaction dataset, these molecules were removed from the reaction entries for the purpose of reaction prediction and enzyme prediction. However, although these molecules are freely available in cell environments, for example NADPH is a coenzyme in anabolic reactions as a reducing agent and it is maintained at a stable concentration in a cell by pentose phosphate pathway for reduction of bio-active molecules,[33] in industrial enzymatic processes, cofactors are difficult to recycle and recover, making cofactor-dependent bio-chemical transformations potentially economically unviable.[34] Therefore, cofactors and free metabolites need to be re-considered when scaling up the biological reactions taking place in a

cell-free bioreactor. A list of cofactors and free metabolites from KEGG database was manually curated by Blaβ *et al*.[35] The full lists can be found in Section 1 of Electronic Supplementary Information (ESI).

Assembly of a biological reaction network is based on graph theory:[36] a network is a mathematical representation of pairwise relationships between objects. In a network G, vertices V representing the objects are connected by edges E. In the case of edges with orientations pointing to one end of the vertices, the graph is defined as directed graph. The mathematical representation of the graph G with m vertices and n edges is shown in Eq. 1 and Eq. 2:

$$G = \left(V(G) = \{v_1, \dots, v_i, \dots, v_m\}, E(G) = \{e_1, \dots, e_j, \dots, e_n\}\right) \quad \text{Eq. 1}$$

$$e_j = \left(v_p, v_q\right), where\ p, q \in (1, \dots, m)\ and\ j \in (1, \dots, n) \quad \text{Eq. 2}$$

Using vertices to represent chemical substances and edges connecting vertices to represent the reactions from reactants to products, the synthetic biological reaction network was generated to include 18,682 molecules, and 10,900 reactions. Since we assumed all metabolic reactions are reversible, the edge in the network have both directions.

Biological reaction prediction

We propose a graph method to predict biological reactions to explore the biochemical reactions domain. The method is target molecule-oriented, *i.e.* for a target molecule of interests t, based on the existing enzymatic transformations in biological reactions, to predict novel reactions from the existing similar functional groups. The diversified reaction space would combine literature-excerpted reactions from the database and calibrate predicted reactions for all target molecules. The workflow of the proposed method is shown in Figure 1.

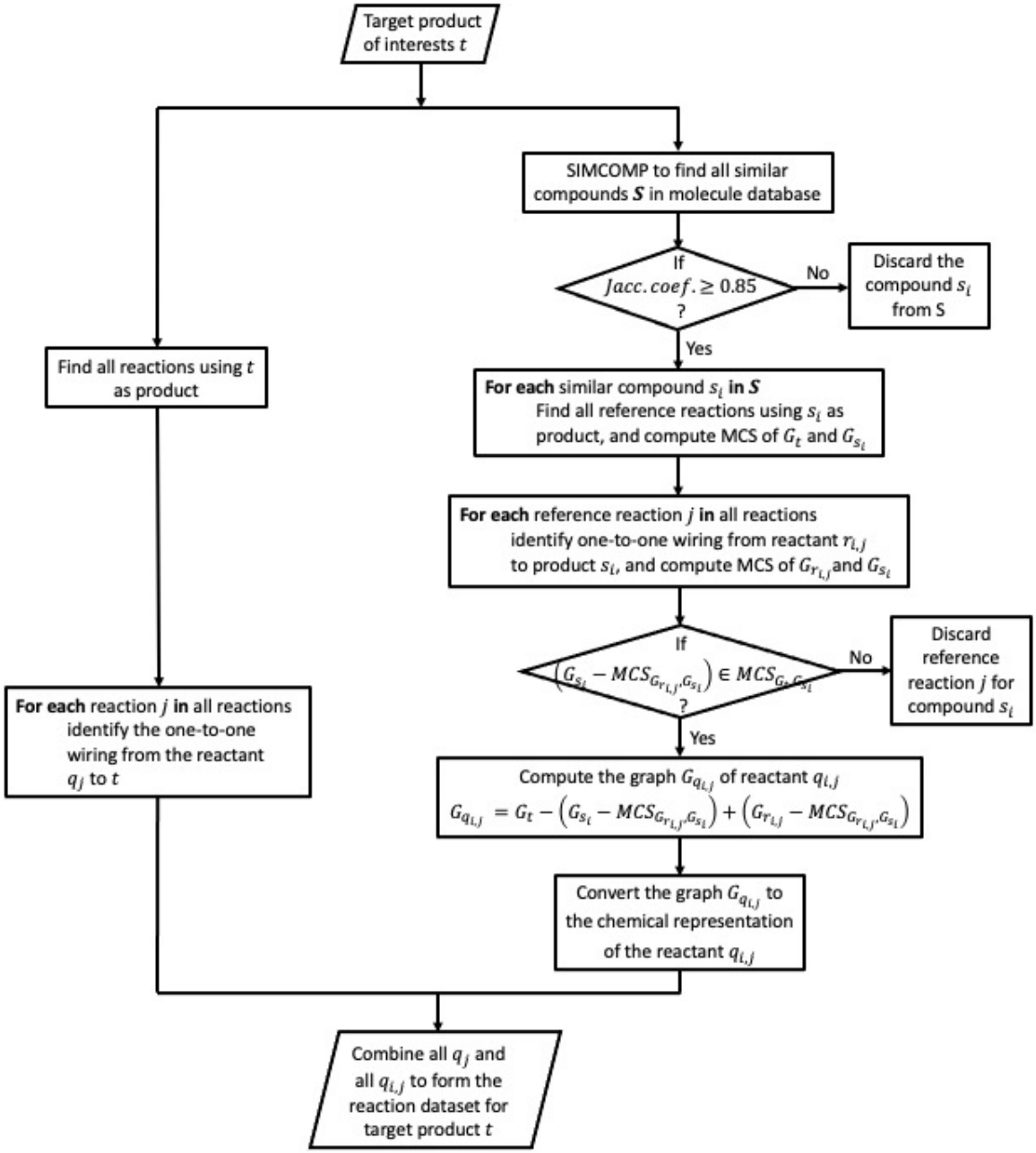


Figure 1. Schematic diagramme of the developed workflow to predict metabolic reactions and diversify the reaction dataset for a target molecule.

For a target molecule t, the algorithm starts by detecting similar compound, s_i, from the KEGG molecule database. To find similar compound, a graph method named SIMilar COMPound (SIMCOMP) designed by Hattori *et al.*[37, 38] was adopted to compare target molecule, t, and all other KEGG molecules. SIMCOMP is believed to successfully cluster the features of

molecules, and especially metabolic molecules.[37] In the method, chemical compounds are converted into two-dimensional undirected graphs, $G \in (V, E)$, where vertices V and edges E represent atoms and covalent bonds respectively. The method excludes all hydrogen atoms and hydrogen attaching covalent bonds. Due to the various atom environments (*i.e.* various attached covalent bonds and adjacent atoms), 68 atom types are defined with numerical codes for all vertex entries: 23 carbon, 18 oxygen, 16 nitrogen, two phosphorus, seven sulphur, four different halogens, and one undefined atom types. For example, saturated carbon atom with three hydrogen and one single-bond functional group attached is defined as 'c1a' carbon atom type, whilst unsaturated carbon atom with a double-bond oxygen and two single bonds attaching to a non-aromatic ring is defined as 'c5x'. The 68 atom types in details are discussed in Hattori *et al*.[37]

By applying SIMCOMP, a dataset of similar compounds, S, is formed for all $s_i \in S$, and a chemical structure similarity indicator Jaccard coefficient is used to retain the compounds, s_i, in S subject to $JC_{t,s_i} \geq 0.85$. Jaccard coefficient, JC_{t,s_i}, between target a molecule t and a similar molecule s_i is determined by Eq. 3, which is the ratio between the intersection set and the union set of the molecular graphs. The above steps were implemented via a Python script to interact with SIMCOMP API.[38] The full list of similar compounds for all KEGG molecules can be found in Section 2 of ESI.

$$JC_{t,s_i} = \frac{G_t \cap G_{s_i}}{G_t \cup G_{s_i}} \qquad \text{Eq. 3}$$

For each detected similar compound $s_i \in S$, all reactions J that produce the compound s_i in KEGG reaction database are found. For each reaction $j \in J$, excluding cofactors and free metabolites from the reaction participants, one-to-one wiring (reactant $r_{i,j}$-to-compound s_i) reactions with only metabolites involved in efficient carbon flux are determined, and defined

as a referenced analogue reaction j. The full list of one-to-one wiring reactions can be found in Section 3 of ESI.

Afterwards, for each analogue reaction $j \in J$ under the level of each detected similar compound $s_i \in S$, the objective is to predict reaction from an unknown reactant $q_{i,j}$ to the target product t based on the functional group transformation at the reaction centre of an analogue reaction j from the reactant $r_{i,j}$ to the similar compound s_i (or in a reversed direction, given that all enzymatic reactions are assumed reversible).

This prediction would be meaningless if the target molecule t does not have the same reaction transformation at the reaction centre, at which the analogue reaction j transforms from the reactant $r_{i,j}$ to the similar compound s_i. To exclude such invalid analogue reactions, graphs $G_{r_{i,j}}$ and G_{s_i} of reactant $r_{i,j}$ and compound s_i are computed respectively, and then maximal common subgraph $MCS_{G_{r_{i,j}}, G_{s_i}}$ is generated for the two graphs. The functional transformation of the analogue reaction takes place in the uncommon area of the reactant $r_{i,j}$ and the similar compound s_i, *i.e.*, in the graph of $\left(G_{s_i} - MCS_{G_{r_{i,j}}, G_{s_i}}\right)$. To ensure the validity of the analogue reaction j, the maximal common subgraph $MCS_{G_t, G_{s_i}}$ between the graphs of the target product t and similar compound s_i is determined, and we check if the functional group is present in the maximal common subgraph $MCS_{G_t, G_{s_i}}$, *i.e.*, if $\left(G_{s_i} - MCS_{G_{r_{i,j}}, G_{s_i}}\right) \in MCS_{G_t, G_{s_i}}$. The failed reactions are removed from the analogue reaction set.

Next, at the graph of target molecule, G_t, the algorithm subtracts the difference between the target product t and the similar compound s_i $\left(G_{s_i} - MCS_{G_{r_{i,j}}, G_{s_i}}\right)$, and then uses the functional

transformation of the analogue reaction $\left(G_{r_{i,j}} - MCS_{G_{r_{i,j}}, G_{s_i}}\right)$ to replace it and to determine the reactant, $q_{i,j}$. Mathematically, the graph of an unknown reactant $G_{q_{i,j}}$ is computed by Eq. 4.

$$G_{q_{i,j}} = G_t - \left(G_{s_i} - MCS_{G_{r_{i,j}}, G_{s_i}}\right) + \left(G_{r_{i,j}} - MCS_{G_{r_{i,j}}, G_{s_i}}\right) \quad \text{Eq. 4}$$

The graph $G_{q_{i,j}}$ is converted to the simplified molecular-input line-entry system (SMILES) string of molecule $q_{i,j}$ for all analogue reactions $j \in J$ under the level of all detected similar compounds $s_i \in S$. Each predicted molecule $q_{i,j}$ and each predicted reaction "$q_{i,j} \leftrightarrow t$" (in the form of one-to-one wiring), are compared with recorded molecules and reactions in database respectively, by their canonical SMILES strings. If the reactions are unique, they are added into the predicted database to diversify the metabolic reaction space.

The workflow was repeated for all KEGG molecules as target molecules t. All predicted metabolic reactions were combined with the literature-excerpted recorded reactions for reaction network assembly. Since this method creates new molecules and reactions, in principle, this method has the potential to predict an infinite number of reactions by iterating the workflow, *i.e.* using created reactions as analogue reactions, and consistently producing novel reactions. However, at this stage, we predicted only from the literature-excerpted reactions (*i.e.* from the first iteration) in KEGG database to avoid uncertainty propagations in further iterations.

The predicted reactions were learned from the reaction rules of analogue reactions, which partially ensures the reactivity. However, similar to most reaction prediction tools,[9-15] the predicted reactions could not be assessed unless they are experimentally screened. Therefore, we added another layer of certainty – pairing predicted reactions with potential promiscuous

enzymes to catalyse the reaction. Suggested by a trained transformer model, predicted reactions with no suitable enzyme to catalyse them were removed from the reaction dataset.

Practically, search of maximal common subgraph between two graphs is a NP-complete problem,[39] which means the computational time increases exponentially when the molecular complexity increases. Therefore, to reduce computational cost, we set a timeout session of five seconds for each search of maximal common subgraph, and discarded the failed search molecules for reaction prediction. We also noticed that SIMCOMP compares similarity of molecules mainly from their two-dimensional features and, therefore, higher order structure differences in molecules (isotopes, stereoisomers and *etc.*) are hard to differentiate by applying SIMCOMP method, whilst enzymatic reactions are stereoisomer specific in most cases. However, this stereoisomer uncertainty could be avoided by applying the enzyme pairing algorithm below – a trained transformer model would learn the stereoisomer specificity from reaction SMILES, and pair with the corresponding enzyme structures, which would remove predicted reactions from infeasible stereoisomers.

Enzyme promiscuity prediction

For the predicted metabolic reactions, enzymes are required to activate and catalyse the metabolic transformation of the substrates. Enzymes usually operate on specific reactions, but however, research shows enzymes are also promiscuous – one enzyme can bind multiple substrates and exhibits broader specificities.[40] An example of promiscuous enzyme malonyl-CoA reductase (MCR)[41] catalysing multiple reactions is shown in Figure 2.

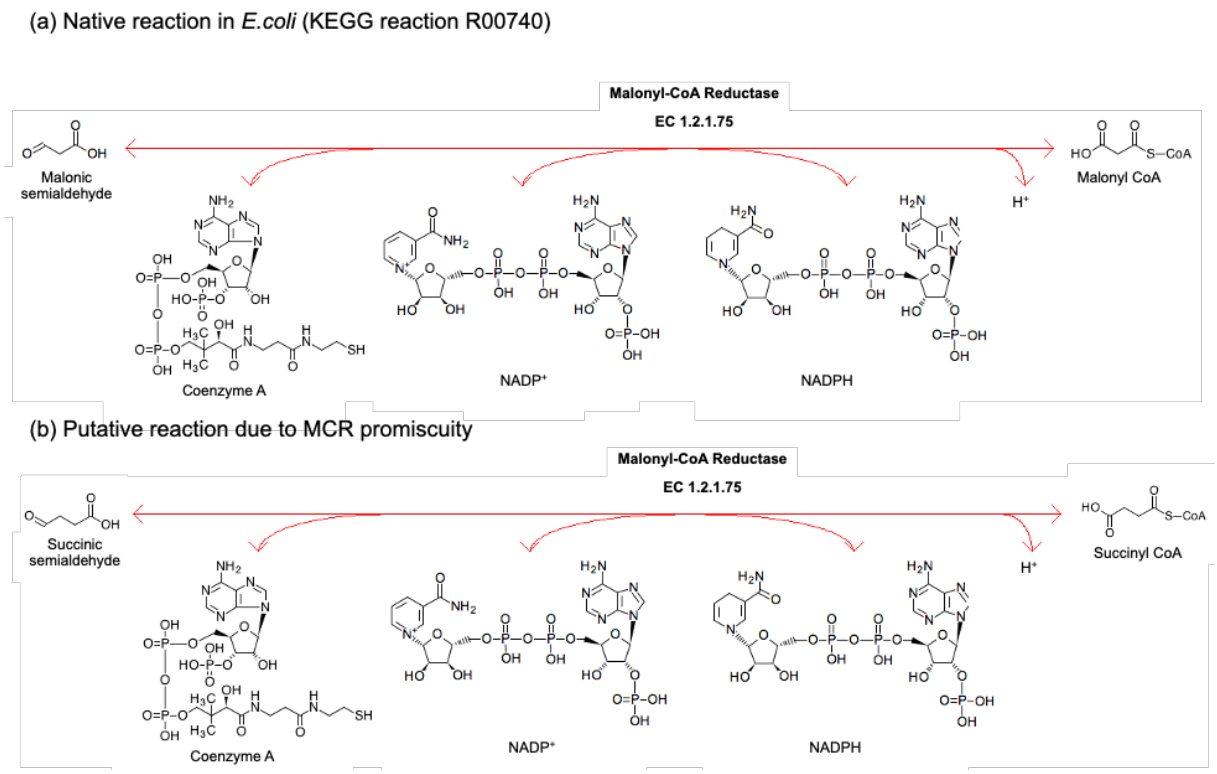


Figure 2 An example of a promiscuous enzyme malonyl-CoA reductase (MCR) (EC number 1.2.1.75) catalysing (a) native reaction of malonyl Coenzyme A reduction (KEGG reaction R00740) in *E.coli* and (b) its putative reaction due to enzyme promiscuity to reduce succinyl Coenzyme A.

We assumed all enzymes were promiscuous and assigned native enzymes to catalyse the predicted reactions with likelihoods by the following the proposed transformer method, inspired from NLP, to translate between a reaction language to an enzyme language.

Reaction language in a transformer

Molecules have been widely converted to machine readable languages such as SMILES,[42] InChI[43] and others. Here we chose SMILES as a language to input into the transformer, since (canonical) SMILES fully specifies the structure of a molecule, and via NLP, SMILES language has been proven successful in tackling multiple chemistry problems.[28, 44] Reaction SMILES

uses defined syntax to combine all reactants, reagents (reaction participants not contributing carbon flow) and products. Reactants, reagents, and products are split by '>' symbol, whilst each molecule in reactants, reagents, and products are split by '.' symbol, see Figure 3a. To input the reaction SMILES strings into transformer, hydrogen atoms and atom mappings were removed from the SMILES strings, and the strings were split/tokenised atom-wise using the regular expression discussed in Schwaller *et al*.,[45] and summarised in Section 4 of ESI.

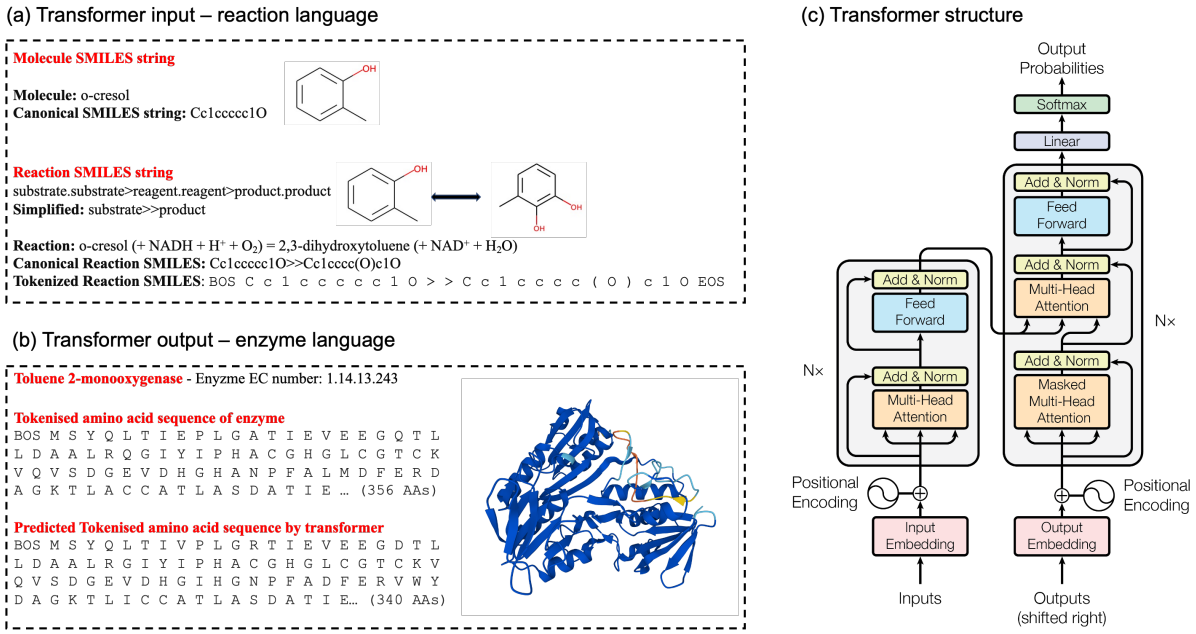


Figure 3. (a) Conversion of reaction into languages of reaction SMILES, with a reaction example of KEGG reaction R03608, catalysed by toluene 2-monooxygenase (EC 1.14.13.243); (b) conversion of enzyme into language of amino acid sequences, schematics of tannase protein structure drawn by AlphaFold2;[20] (c) schematics of transformer model structure, reproduced from Vaswani *et al*.[26]

Enzyme language in a transformer

Enzymes in KEGG database are recorded with their enzyme commission (EC) number, which fully specifies the organisms that host the enzyme and the specific enzymatic transformations, and uses a topological classification scheme to categorise enzymes/enzymatic transformations

into four levels.[46] For example, from Figure 2, the EC number of enzyme MCR is 1.2.1.75, where '1' at the first level represents the main category of oxidation/reduction reaction, '2' at the second level represents the reaction centre of an aldehyde or carbonyl group, '1' at the third level represents the acceptors of NAD^+ or $NADP^+$, '75' at the fourth level represents the malonate semialdehyde forming reaction specifically. The categories of reaction types of the first EC level are summarised in Section 5 of ESI.

The binding between substrates and enzymes is a function of the secondary and tertiary structures of proteins, *i.e.* the physiochemical property of constituting amino acids to form the peptide chains of the proteins.[21] We crawled genes which transcribe and translate into the protein of the enzymes from KEGG genes database for all recorded enzymes. KEGG genes database also gives the amino acid sequences of the specific gene. One enzyme usually links with multiple genes, varied from the cell organism (e.g. human serum albumin (hsa), bacterial, *etc.*), which originally hosts this enzyme. However, the amino acid sequences of the different genes are usually close. For example, for MCR – EC 1.2.1.75, two genes are reported: (1) mse: Msed_0709, and (2) sto: STK_21710, whilst both genes are linked with identical 357 amino acid sequences. For each enzyme, we only included one gene/amino acid sequence to include in the dataset. To minimise the biases on the selection from multiple amino acid options to affect the model results, the principles to select the gene/amino acid were: to (1) select the reported literature sources of the enzyme which were mostly cited, (2) stick to one cell organism if possible – hsa. The selected gene names and amino acid sequences chosen for the model are given in Section 6 of ESI.

A peptide chain of a protein usually contains 21 types of amino acids, varied from the constituting side chain functional groups. 'X' denotes unknown amino acid in the peptide. Here

we use 22 letters of amino acids to denote the amino acid sequence of enzymes as an enzyme language for the transformer model. The specification between the letters and amino acids is given in Section 6 of ESI. No segmentation is required for the amino acid sequence – the sequence is tokenized letter-wise into the transformer model, see Figure 3b.

Data Pre-processing

In this method, a transformer model was trained to predict the correct mapping between enzymes and their binding substrate structures. To do this, we collected all one (substrate)-to-one (product) reactions from KEGG. Note that in this context, reagents are explicitly excluded from substrates/products. For example, for the reaction in Figure 2a, NADPH, hydrogen cation and NADP+ are regarded as reagents. Also, a great portion of KEGG reactions could not produce valid reaction SMILES strings. After filtering these invalid reactions, eventually 3,079 recorded enzymes and 3,594 linked reactions are included as one-to-one reactions, details shown in Section 3 of ESI. This indicates nearly 500 reactions in this reaction set are catalysed by promiscuous enzymes.

For all molecules in the reactions, SMILES strings were canonicalised by RDKit to assemble the reaction SMILES. To increase the number of reactions in the dataset, for each reaction, all molecular canonical SMILES were randomised three times by RDKit to reassemble three different reaction SMILES strings. These are reactions containing identical transformations but different representations. By adding the randomised reaction SMILES strings, the reaction datapoints were quadrupled, and eventually 3,079 enzymes are linked with 14,376 reactions.

The model could not use the full length of all amino acid sequences (maximum length of 1,411), since this would significantly increase the requirement for GPUs. Therefore, the amino acid

sequences were padded. Most peptide chains of the enzymes have the lengths shorter than 400, where the statistics of the peptide chains lengths are shown in Figure S1, ESI. The average length of amino acid sequences is 449, whilst the median is 395. The length taken into the model input, 600, covers the length of up to 78% of the peptide chains, which is reasonable to retain most information. In the enzyme language, we left-padded the first 600 letters of the amino acid, and used zero to denote the vacancies from the right-hand side for short sequences.

For both the reaction and enzyme languages, two extra tokens were added to each string – 'BOS' and 'EOS', which are the start and finish signals of the two strings to the machine. To translate the reaction and enzyme languages into the transformer model, the tokens in both languages were converted into numerical indices, based on the one-to-one token-numerical index dictionaries of the two languages, shown in Section 4 of ESI. The vectors of numerical indices were then able to be processed into the model.

Transformer model structure

The transformer model is a deep learning neural network that differentiates the weights of each part of the input by adopting the mechanism of self-attention.[26] It includes the building blocks of input/output embedding, positional encoding, encoders and decoders. Details of model structure can be found in Vaswani *et al*.[26] and are summarised below. Schematics of the transformer model structure are shown in Figure 3c.

The input/output embedding layer intakes token indices in the input/output languages, and against each of those indices, an embedding vector is attached. These vectors are initially filled with random numbers, and these values are updated while training to capture the intrinsic linguistic features of the input/output tokens/indices. Since all tokens/indices in a sentence (*i.e.*

the reaction SMILES and amino acid sequences) are passed into the embedding layer at once, the model is not aware of the sequential information of the sentence until positional encoding is applied to the embedding vector of each index to capture position by wave frequencies, commonly computed by Eq. 5.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad \text{Eq. 5}$$

In Eq. 5, pos is the position of the index in the sentence, i is the index of each of the position embedding dimensions, and d is the dimension of the embedding layer.

The position encoding of the indices embedding is then fed into multiple encoder and decoder units. Each encoder unit consists of a multi-head attention sub-unit and a feed forward sub-unit. The multi-head attention sub-unit consists of numbers of parallel scaled-dot attention layers to compute three input matrices – keys K, values V, and queries Q, and determine the attention of each index by Eq. 6, where d_k is the dimension of the key vector K to scale the dot product QK^T.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{Eq. 6}$$

In Eq. 6, the dot product QK^T determines the closeness of the keys aligned with the queries. A value vector is associated with each key, where the value is multiplied by the softmax transformation of the scaled dot product, through which the dot product is normalised, and the large components are emphasised. The feed forward unit consists of numbers of deep learning neural network layers and computes features of indices, which are then queried by the next encoder unit or a decoder unit.

As shown in Figure 3c, a decoder unit is similar to the encoder, but an additional encoder-decoder attention sub-unit is inserted between the multi-head attention sub-unit and the feed forward sub-unit to add related information from features of indices computed from the encoder. After a generator unit consisting of a linear layer and a softmax layer, the model eventually predicts the amino acid sequence, and gives probability distribution between the actual amino acid sequence and the predicted amino acid sequence based on the loss function. The transformer model was implemented in PyTorch package at the platform of Google Colab.

Results and Discussion

Reaction prediction results and visualisation

All predicted reactions and all recorded reactions producing the target product *t* are combined into the local database. From the 18,682 molecules and 3,079 single-step reactions parsed from KEGG database, with one iteration of the reaction prediction workflow, 32,990 reactions were predicted, among which 20,568 reactions were either redundant (predicted reactions being previously reported within KEGG or Reaxys records), or not able to use the proposed method to find suitable enzymes to catalyse. After removing the redundant and invalid reactions, 12,422 predicted reactions were added into the local reaction database.

The inferred transformations were applied to known substrates, and reaction product molecules were inferred with given canonical SMILES strings. By comparing these predicted reaction products with existing molecules in KEGG and Reaxys datasets, it is shown that 7,827 molecules were not found in either. We know that the molecule datasets that we compiled from KEGG and Reaxys are not comprehensive in the coverage of all known molecular space. Some of these molecules could be found in other knowledge bases, such as CAS[47] and PubChem,[48]

or elsewhere in literature and patents. These reaction products from predicted reactions might include some molecules never reported, but they were not *de novo* designed molecules, *i.e.* the rationally designed molecular structure based on their functionalities.[49] The purpose for this step was not to design new molecules, but apply generalised bio-catalytic transformations to existing metabolic substrates. This may lead to some unknown products or intermediates, having potential to be used for drug discovery, or to synthesise some target drug molecules in the future. Here we do not claim that this method is recommending new products / substrates with a high degree of confidence in their practical usefulness or synthesizability.

An example of predicted reactions is shown in Figure 4a, and the atoms of main reactant and product of the reactions are mapped and shown in Figure 4b, by a chemically agnostic attention-guided reaction mapper, namely, RXNMapper.[50] The analogue reaction is the glycosylation of 4-hydroxymandelonitrile, catalysed by hydroxymandelonitrile glucosyltransferase, to produce taxiphyllin, where the glycosyl group in taxiphyllin is transferred from cofactor UDP-glucose. The reaction centre of the analogue reaction is an oxygen atom, marked as "O:4" at the analogue reaction atom mapping. Similarly, the predicted reaction is the glycosylation of mandelonitrile to produce prunasin. The reaction centre and glycosyl group transformation remains identical as the analogue reaction, except both the predicted reactant and product are lack of a hydroxyl group, marked as 'O:20H' at atom mapping, attached to the benzene ring. Although it is suspected the electron donation of the conjugated hydroxyl group "O:20H" might slightly affect the adjacent atom environment near the reaction centre, the original publication[51] of the analogue reaction suggests that in this transformations, most functional groups could possibly substitute the phenyl group attached to the carbon atom "C:3", for example, a 3,4-dihydroxylphenyl group (*i.e.* the addition of a hydroxyl group on the carbon atom "C:18"). This enhances the feasibility of the predicted reaction.

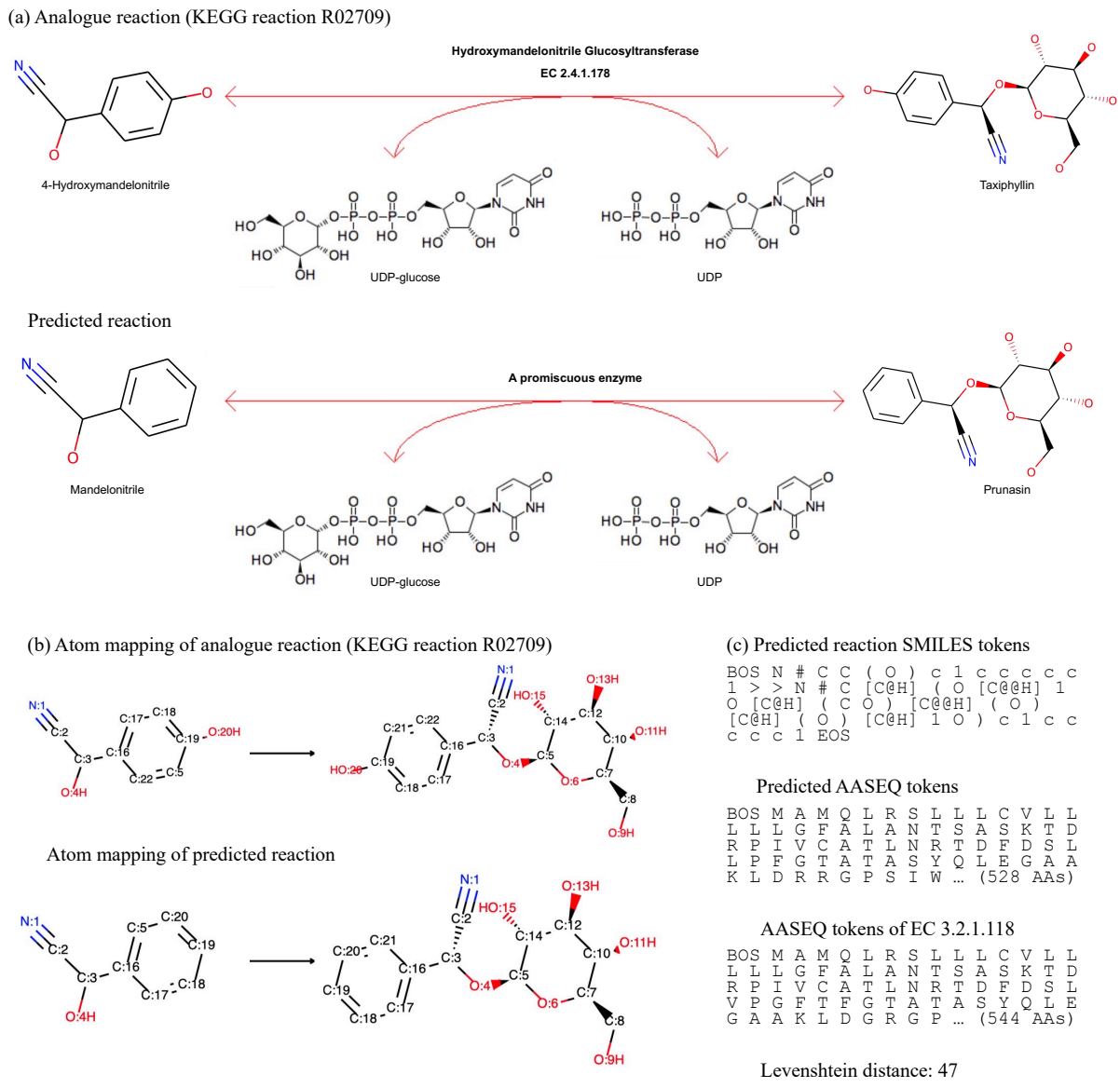


Figure 4. An example of a reaction predicted by the workflow described in Figure 1: (a) schematics of the analogue reaction, glycosylation of 4-hydroxymandelonitrile, and the predicted reaction, glycosylation of mandelonitrile; (b) atom mapping of the main reactant and main product of the reactions by RXNMapper;[50] (c) the predicted amino acid sequence of the catalyst for the predicted reaction by the trained transformer.

The biological reaction space appears to have a complicated structure. To understand the comprehension of the biological reaction space with the addition of the predicted reactions, biological reactions were visualised by a t-distributed stochastic neighbour embedding (t-SNE)

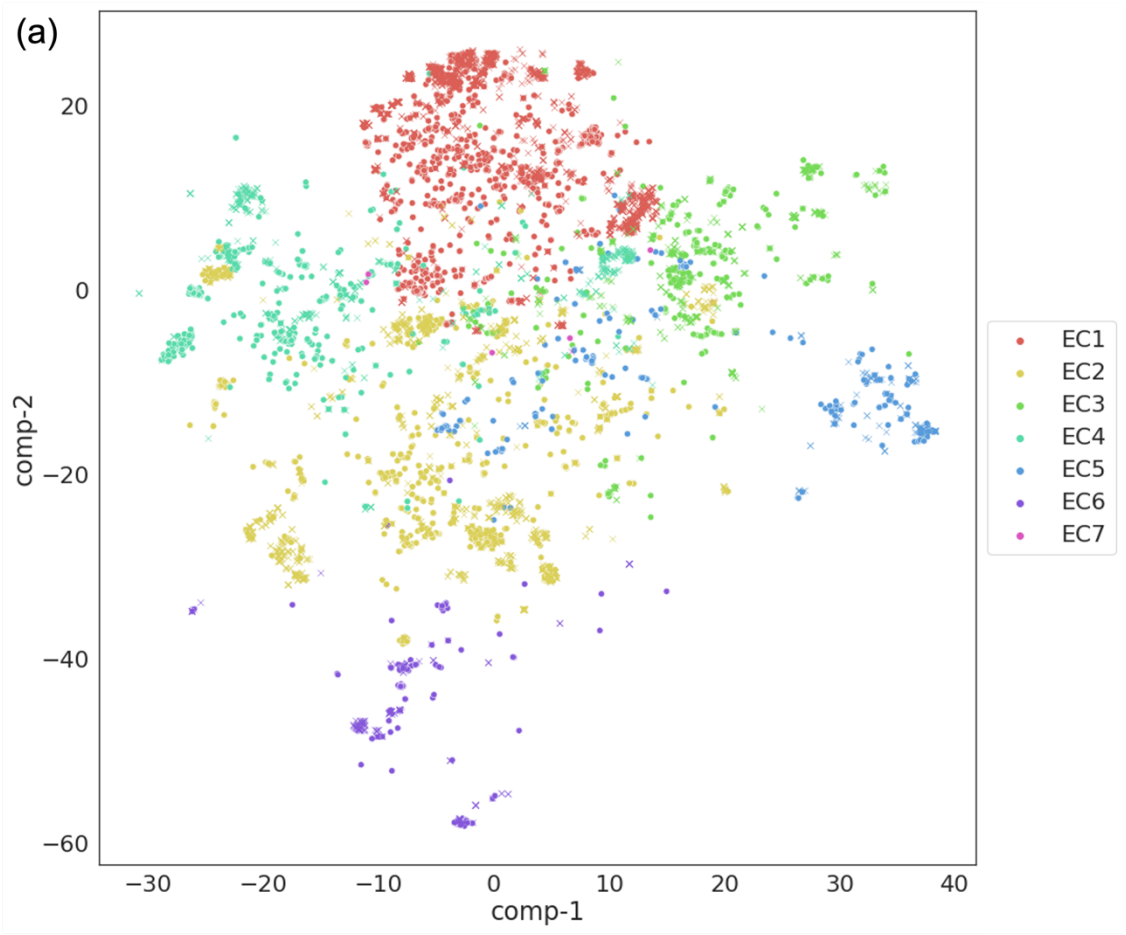
diagram,[52] which is a non-linear dimensionality reduction technique to convert similarities between data to joint probabilities and minimise the KL divergence (Eq. 7) between the probability distribution. t-SNE locates reactions into lower dimensional map. Here we chose to visualise reactions in two dimensions.

To encode molecules and reactions for visualisation, we evaluated two reaction fingerprints to convert molecular structures into numerical representations, which are extended-connectivity fingerprints (ECFP),[53] and reaction bidirectional encoder representations from transformers fingerprint (BERT FP).[54] ECFP is a topological molecular fingerprint to convert the circular structure of neighbourhood of each non-hydrogen atom into bytes. In this work, the radius of the fingerprint was three (ECFP3), which detected the multiple layers of the neighbourhoods from the molecule centre, and all molecules were converted into 256 fixed-length bit string to reduce the chance for bit collision. To represent the enzymatic transformation via a reaction, the ECFP3 of reaction product was subtracted with that of substrate to create 256 bits as the reaction fingerprint. BERT FP uses a trained BERT transformer model,[54] learned from USPTO reaction dataset,[55] which covers comprehensive reaction types, to encode reaction SMILES into 256 numerical values. The KEGG curated reactions were grouped by the first level of corresponding enzyme EC numbers, whilst the predicted reactions were grouped by that of most recommended enzyme from transformer.

For t-SNE implementation, parameters were set after tuning: perplexity, the number of nearest neighbours in the algorithm was set to 30, early exaggeration, which controls the tightness of clusters in the space, was set to 12, and learning rate was set to 100.

After 1,000 iterations of computing t-SNE from the combined curated reactions and predicted reactions dataset, the KL divergences (Eq. 7) for ECFP and BERT FP are 2.58 and 0.82 respectively. BERT FP better clusters the metabolic reactions space, where the t-SNE biological space is shown in Figure 5. That by ECFP, shown in Figure S2 in ESI, resembling a 'ball' with points approximately equidistant from its nearest neighbours, indicates the failure of digitalisation of reactions into ECFP fingerprints. Figure 5a shows most reactions, including curated and predicted reactions are well separated at different regions in the two-dimensional space, except EC1 and EC2 reactions are partially dispersed into other regions. This is likely due to the large coverages of the EC1 and EC2 reactions, which cover oxidation/reduction reactions, and reactions transferring functional groups respectively. These reactions have possible intersections with other reaction types. Moreover, most predicted reactions are located near the regions of their reaction clusters. This assures the feasibilities of predicted reactions.

To better understand the biochemical space, EC1 reactions were zoomed in and clustered based on the second level of EC numbers, as shown in Figure 5b. Figure 5b shows that most reaction types are still able to be separated into different clusters but with fusion into other reaction types. This indicates the BERT FP and t-SNE combined techniques cluster reactions, but are not able to interpret reactions into specific levels. Two examples of gathering scatters are shown in Figure 5b, which interprets the common rules of the reactions in the corresponding gathering scatters – red dots and green crosses: oxidation, and esterification of alcohols respectively, falling into the reaction categories of EC1.1 and EC1.8.



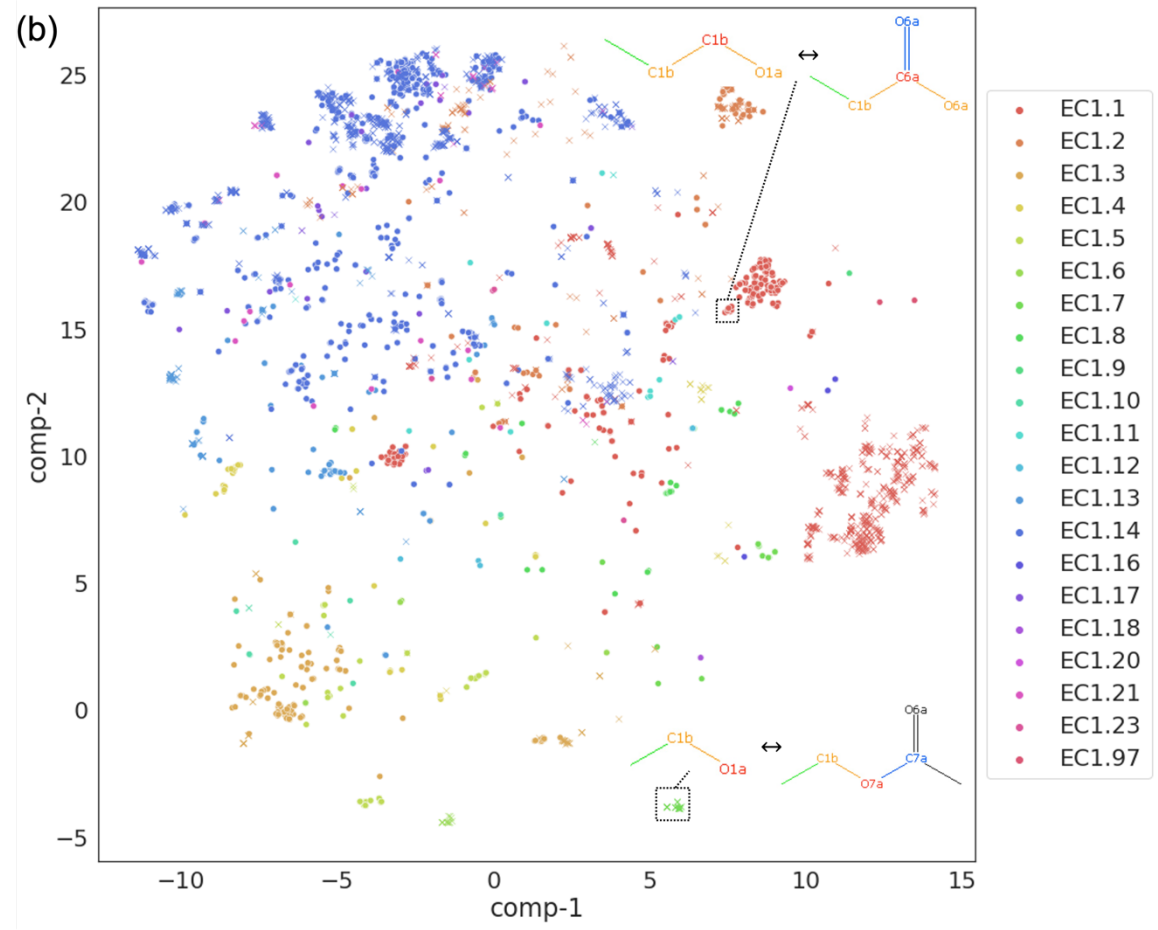


Figure 5. Two-dimensional t-SNE diagram to visualise biochemical reaction space from KEGG curated reactions and predicted reactions using BERT fingerprint. All KEGG curated reactions are shown with dot (·), whilst predicted reactions are shown with cross (×). (a) Reactions clustered by the first level of EC numbers; (b) 'zoom in' of EC1 reactions clustered by the second level of EC numbers, with two examples interpreting reaction rules of gathering scatters. The codes (for example, 'c1b') inside molecules indicate atom environments defined by Hattori *et al*.[37]

Enzyme promiscuity prediction

The enzyme promiscuity prediction assigns most suitable enzymes in the database to predicted reactions based on a trained transformer model. The model and prediction results are shown below.

Transformer model training and assessment

The reaction – enzyme dataset (quadrupled by randomised SMILES technique, as described above) was split into training, validation and test datasets by the ratio of 7:2:1. Since the reaction – enzyme pairs were not evenly distributed in terms of its enzyme EC categories (first level of EC number), the dataset was deliberately split by its EC categories to include same ratio of reaction types into each of training, validation and test datasets. A loss function of pairwise Kullback-Leibler (KL) divergence[56] was used to measure the probability distribution of the true output language encoding y_{true} and computed one y_{pred}, as shown in Eq. 7.

$$KL(y_{pred}, y_{true}) = y_{true} \cdot \log \frac{y_{true}}{y_{pred}} \qquad \text{Eq. 7}$$

Comparison of validation losses was used to tune the hyperparameters for the transformer model, where the training and validation results for different hyperparameter settings are shown in Figure 6.

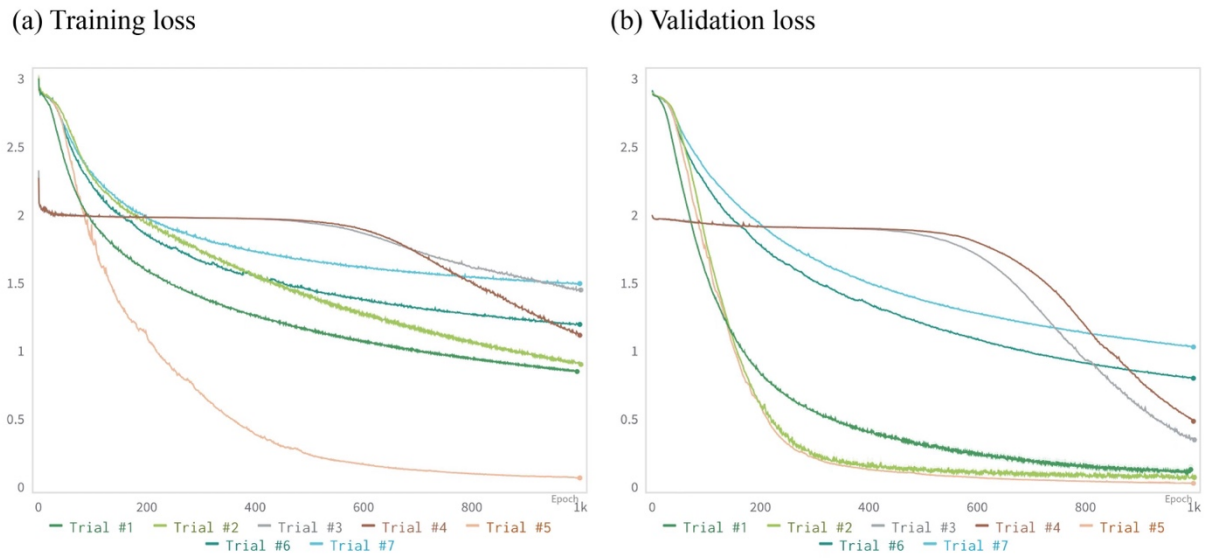


Figure 6. (a) Training and (b) validation KL divergence losses of the transformer model based on different parameter settings. The details and results of model settings for the training trials are shown in Table S2, ESI.

From Figure 6 and Table S2, it is noticed that in most trials, validation loss is consistently lower than training loss during training epochs. Some of them are due to the use of dropout, where dropout randomly freezes neurons to penalise model variance during model training, and this only has effects on the training loss. This is also possibly due to the manual split of training and validation data based on EC reaction types, where the validation data is possibly naturally less noisy.

The trials #3 and #4 attempted cross entropy as loss function, which both have slight descents in the beginning and sudden drops after 500 epochs. The model parameters learned slowly due to the use of cross entropy, which fell into certain intervals in the beginning, and were only able

to learn properly from 500 epochs. The final validation loss after 2000 epochs for trails #3 and #4 are 0.28, and 0.25 respectively. However, for the convenience of visualisation, only first 1000 epochs are shown in Figure 6.

The final settings for the transformer were chosen from the lowest validation KL divergence loss after 1,000 or 2,000 epochs, where some trials were early terminated to save computational costs when training and validation losses become stable. Dimension of embedding for the indices was 128. Six encoder and decoder units were assembled. In each encoder and decoder unit, the number of hidden neurons for multi-head attention was four, and there are four hidden layers and the dimension of the feed forward unit in each layer was 128 with dropout rate of 0.1. Adam optimiser was used to update model parameters with a varied learning rate introduced from Vaswani *et al*.[26] The batch size was 16 to minimise GPU memory requirement (the length of enzyme sentence could reach max 600 indices, which converts to a large tensor after encoding). The training loss and validation loss for the final trained transformer model described above are 0.068 and 0.029 respectively, and the model was assessed by the KL divergence loss of the test dataset, 0.035, which proves the model accuracy at unseen datapoints. It is noticed that with the current model setting, the number of trainable parameters in the model is 343,064, with 200,064 from the encoders, 267,136 from the decoders, 6,400 from the embeddings, and 1,560 from the final generator, whilst the total number of training data is 10,063 (70% of 14,376 reactions, quadrupled from randomised reaction SMILES). However, it is normal in a transformer model to have training data much less than trainable parameters, for example the Molecular Transformer with 65 million trainable parameters and 0.48 million datapoints.[44]

The predicted amino acid sequences are compared with the predicted amino acid sequence of enzyme by Levenshtein distance.[57] The algorithm of Levenshtein distance is described in Eq. 8, which determines the minimum steps of edits between two strings, and by edits, it means the insertion, deletion or substitution of a single character in a string. In Eq. 8, the Levenshtein distance $lev_{a,b}$ between two strings a and b is a function of their terminal character postion i and j, respectively.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0, \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & otherwise. \end{cases} \quad \text{Eq. 8}$$

An example of the final transformer model result from the test dataset is shown in Figure 3b, which compares the predicted amino acid sequence and the actual amino acid sequence for tannase (EC 1.14.13.243), catalysing KEGG reaction R03608 of oxidation of o-cresol to 2,3-dihydroxytoluene. The predicted amino acid has longer length than the actual one, which are 340 and 356 respectively, and the Levenshtein distance between the two sequences are 66. The comparison between all predicted amino acid sequence and the actual amino acid are shown in Section 9 of ESI. The mean and median of Levenshtein distance between the amino acids are 64.6 and 62 respectively, which indicates in average, approximately 65 edits are required to convert a predicted amino acid sequence into the actual one, where the average amino acid sequence length is 414.

Transformer model prediction

To pair a predicted reaction with possible enzymes, a trained transformer computes the predicted amino acid sequences, which is compared with all available enzymes in the pool.

A Levenshtein distance threshold of 64.6 - the mean value in test dataset, between the predicted sequence and available sequences was used to measure the possibility of enzyme promiscuity. For a given predicted reaction, possible promiscuous enzymes below this threshold were sorted by Levenshtein distance to rank the likelihood to catalyse the predicted reaction. Reactions with no enzyme suggested were removed from the predicted reaction dataset. The transformer model prediction for the predicted reaction example of glycosylation of mandelonitrile is shown in Figure 4c. From the final trained model, an amino acid sequence with length of 528 is predicted for the predicted reaction SMILES. Using the Levenshtein distance to filter the invalid enzymes, only enzyme prunasin hydrolase (EC 3.2.1.118) falls into the possible promiscuous enzyme list, with Levenshtein distance of 47. Prunasin hydrolase originally also catalyses the glycosylation of mandelonitrile, but with a cofactor D-glucose. This suggests that, with an identical substrate but slightly putative reaction mechanism, the enzyme could possibly catalyse the predicted reaction.

Although the transformer model gives a specific amino acid sequence for a predicted enzyme, at this stage it is not suggested to directly evolve an enzyme based on the exact sequence. This is due to the model prediction uncertainty, which causes the distance of the predicted amino acid sequence from ground truth. To the best of our knowledge, this could only be possible with the expansion of synthetic biological reaction data to a much larger size, where a more comprehensive transformer could be trained to identify the intrinsic mechanisms between reactions and amino acid sequences by attention units of the transformer. However, by comparing the Levenshtein distance, this novel approach helps to: (1) calibrate the feasibility of predicted reactions, and (2) reduce the search scope for promiscuous enzymes in a pool, and therefore increase the efficiency of exploration of the biochemical reactions space.

Conclusions

A workflow was developed to explore the sparse domain of biochemical transformations and extends on our previous work[5] of hybrid chemical and biological reaction networks. Literature-excerpted biological reaction data recorded in KEGG database was mined. The biochemical reactions database size (specifically, the one substrate-to-one product wiring reactions) was amplified four times by predicting possible biochemical reactions from the KEGG reactions. From a molecular graph method, a reaction centre was identified for an analogue reaction, and the functional transformation at the reaction centre was suggested to valid similar compounds to predict not reported reactions.

To catalyse the novel reactions, instead of designing new enzymes, which was believed to be uncertain and time-consuming, we focused on enzymatic promiscuity, which expands the specificity of the native enzymes to putative substrates. A deep learning transformer model translating the languages of reaction SMILES and enzyme amino acid sequences of enzymes was trained to learn from the reaction transformations and the protein structures of enzymes, and subsequently suggest on promiscuous enzymes to bind with the substrates in the predicted reactions. The proposed transformer model helped calibrate the feasibility of the predicted reactions and reduce the search scope for promiscuous enzymes in the pool. Eventually, 12,422 novel reactions were predicted, and promiscuous enzymes were suggested to increase the confidence to synthesise these reactions. The populated biological space was also visualised by t-SNE to understand reaction clustering.

For future work, the populated biological space would be merged into the hybrid organic chemical and synthetic biological reaction network, and we would investigate the added values

of the populated biological space to guide synthetic route planning of valuable pharmaceutical molecules.

Acknowledgements

CZ is grateful to Cambridge Trust CSC Scholarship for funding their PhD study. We gratefully acknowledge collaboration with RELX Intellectual Properties SA and their technical support, which enabled us to mine Reaxys. Copyright © 2020 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited. Reaxys data were made accessible to our research project via the Elsevier R&D Collaboration Network. This work was in part supported by National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program as a part of the Cambridge Centre for Advanced Research and Education in Singapore Ltd (CARES).

Data Availability

KEGG reaction and molecule data are available via KEGG APIs. All other data are shared via Supporting Information.

Conflicts of Interest

Authors do not have any conflicts of interests to report.

References

1. Ko, Y.-S.; Kim, J. W.; Lee, J. A.; Han, T.; Kim, G. B.; Park, J. E.; Lee, S. Y., Tools and strategies of systems metabolic engineering for the development of microbial cell factories for chemical production. *Chem. Soc. Rev.* **2020,** *49* (14), 4615-4636.
2. Lee, S. Y.; Kim, H. U.; Chae, T. U.; Cho, J. S.; Kim, J. W.; Shin, J. H.; Kim, D. I.; Ko, Y.-S.; Jang, W. D.; Jang, Y.-S., A comprehensive metabolic map for production of bio-based chemicals. *Nat. Catal.* **2019,** *2* (1), 18-33.
3. Weber, J. M.; Guo, Z.; Zhang, C.; Schweidtmann, A. M.; Lapkin, A. A., Chemical data intelligence for sustainable chemistry. *Chem. Soc. Rev.* **2021,** *50* (21), 12013-12036.
4. Sheldon, R. A.; Woodley, J. M., Role of Biocatalysis in Sustainable Chemistry. *Chemical Reviews* **2018,** *118* (2), 801-838.

5. Zhang, C.; Lapkin, A. A., Hybridizing Organic Chemistry and Synthetic Biology Reaction Networks for Optimizing Synthesis Routes. *ChemRxiv* **2022**.
6. Wang, L.; Dash, S.; Ng, C. Y.; Maranas, C. D., A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst.* **2017,** *2* (4), 243-252.
7. Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J., RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **2021,** *4* (2), 98-104.
8. Levin, I.; Liu, M.; Voigt, C. A.; Coley, C. W., Merging enzymatic and synthetic chemistry with computational synthesis planning. *Nature Communications* **2022,** *13* (1), 7747.
9. Sivakumar, T. V.; Giri, V.; Park, J. H.; Kim, T. Y.; Bhaduri, A., ReactPRED: a tool to predict and analyze biochemical reactions. *Bioinform.* **2016,** *32* (22), 3522-3524.
10. Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S., BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminformatics* **2019,** *11* (1), 2.
11. Ridder, L.; Wagener, M., SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **2008,** *3* (5), 821-832.
12. de Bruyn Kops, C.; Šícho, M.; Mazzolari, A.; Kirchmair, J., GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics. *Chem. Res. Toxicol.* **2021,** *34* (2), 286-299.
13. Litsa, E. E.; Das, P.; Kavraki, L. E., Machine learning models in the prediction of drug metabolism: challenges and future perspectives. *Expert Opin. Drug Metab. Toxicol.* **2021**, 1-3.
14. Litsa, E. E.; Das, P.; Kavraki, L. E., Prediction of drug metabolites using neural machine translation. *Chem. Sci.* **2020,** *11* (47), 12777-12788.
15. Jiang, J.; Liu, L.-P.; Hassoun, S., Learning graph representations of biochemical networks and its application to enzymatic link prediction. *Bioinform.* **2021,** *37* (6), 793-799.
16. Richardson, J. S.; Richardson, D. C., The de novo design of protein structures. *Trends Biochem. Sci.* **1989,** *14* (7), 304-309.
17. Cobb, R. E.; Chao, R.; Zhao, H., Directed evolution: Past, present, and future. *AIChE J.* **2013,** *59* (5), 1432-1440.
18. Lutz, S., Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* **2010,** *21* (6), 734-743.
19. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D., Improved protein structure prediction using potentials from deep learning. *Nature* **2020,** *577* (7792), 706-710.
20. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583-589.
21. Mazurenko, S.; Prokop, Z.; Damborsky, J., Machine Learning in Enzyme Engineering. *ACS Catal.* **2020,** *10* (2), 1210-1223.
22. Hult, K.; Berglund, P., Enzyme promiscuity: mechanism and applications. *Trends Biotechnol.* **2007,** *25* (5), 231-238.

23. Tawfik, O. K.; Dan, S., Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry* **2010,** *79* (1), 471-505.
24. Almonacid, D. E.; Babbitt, P. C., Toward mechanistic classification of enzyme functions. *Current Opinion in Chemical Biology* **2011,** *15* (3), 435-442.
25. Hochreiter, S.; Schmidhuber, J., Long Short-Term Memory. *Neural Computation* **1997,** *9* (8), 1735-1780.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I., Attention Is All You Need. In *Adv. Neural Inf. Process Syst.*, 2017; pp 5998-6008.
27. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics.: Minneapolis, Minnesota, 2019; Vol. 1, pp 4171–4186.
28. Kreutter, D.; Schwaller, P.; Reymond, J.-L., Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **2021,** *12* (25), 8648-8659.
29. Probst, D.; Manica, M.; Nana Teukam, Y. G.; Castrogiovanni, A.; Paratore, F.; Laino, T., Biocatalysed synthesis planning using data-driven learning. *Nature Communications* **2022,** *13* (1), 964.
30. Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T., Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **2020,** *11* (12), 3316-3325.
31. Ofer, D.; Brandes, N.; Linial, M., The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal* **2021,** *19*, 1750-1758.
32. Kanehisa, M.; Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000,** *28* (1), 27-30.
33. Alfarouk, K. O.; Ahmed, S. B. M.; Elliott, R. L.; Benoit, A.; Alqahtani, S. S.; Ibrahim, M. E.; Bashir, A. H. H.; Alhoufie, S. T. S.; Elhassan, G. O.; Wales, C. C.; Schwartz, L. H.; Ali, H. S.; Ahmed, A.; Forde, P. F.; Devesa, J.; Cardone, R. A.; Fais, S.; Harguindey, S.; Reshkin, S. J., The Pentose Phosphate Pathway Dynamics in Cancer and Its Dependency on Intracellular pH. *Metabolites* **2020,** *10* (7).
34. Bowie, J. U.; Sherkhanov, S.; Korman, T. P.; Valliere, M. A.; Opgenorth, P. H.; Liu, H., Synthetic Biochemistry: The Bio-inspired Cell-Free Approach to Commodity Chemical Production. *Trends Biotechnol.* **2020,** *38* (7), 766-778.
35. Blaß, L. K.; Weyler, C.; Heinzle, E., Network design and analysis for multi-enzyme biocatalysis. *BMC Bioinform.* **2017,** *18* (1), 366.
36. West, D. B., *Introduction to graph theory*. Second Edition ed.; Pearson: United States, 2018.
37. Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M., Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **2003,** *125* (39), 11853-65.
38. Hattori, M.; Tanaka, N.; Kanehisa, M.; Goto, S., SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.* **2010,** *38* (suppl_2), W652-W656.
39. Raymond, J. W.; Willett, P., Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.* **2002,** *16* (7), 521-533.
40. Baas, B. J.; Zandvoort, E.; Geertsema, E. M.; Poelarends, G. J., Recent advances in the study of enzyme promiscuity in the tautomerase superfamily. *Chembiochem* **2013,** *14* (8), 917-26.

41. Kockelkorn, D.; Fuchs, G., Malonic semialdehyde reductase, succinic semialdehyde reductase, and succinyl-coenzyme A reductase from Metallosphaera sedula: enzymes of the autotrophic 3-hydroxypropionate/4-hydroxybutyrate cycle in Sulfolobales. *J Bacteriol* **2009,** *191* (20), 6352-6362.
42. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988,** *28* (1), 31-36.
43. Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D., InChI, the IUPAC International Chemical Identifier. *J. Cheminformatics* **2015,** *7* (1), 23.
44. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A., Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019,** *5* (9), 1572-1583.
45. Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T., "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018,** *9* (28), 6091-6098.
46. Enzyme nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Pp 862. Academic Press, San Diego. 1992 ISBN 0-12-227165-3. *Biochemical Education* **1993,** *21* (2), 102-102.
47. CAS Registry System. *Journal of Chemical Information and Computer Sciences* **1978,** *18* (1), 58-58.
48. PubChem Explore Chemistry. https://pubchem.ncbi.nlm.nih.gov (accessed 15 Nov).
49. Meyers, J.; Fabian, B.; Brown, N., De novo molecular design and generative models. *Drug Discovery Today* **2021,** *26* (11), 2707-2715.
50. Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T., Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances 7* (15), eabe4166.
51. Hösel, W.; Schiel, O., Biosynthesis of cyanogenic glucosides: In vitro analysis of the glucosylation step. *Archives of Biochemistry and Biophysics* **1984,** *229* (1), 177-186.
52. Maaten, L. v. d.; Hinton, G., Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008,** *9* (86), 2579-2605.
53. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010,** *50* (5), 742-754.
54. Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L., Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **2021,** *3* (2), 144-152.
55. Lowe, D. Chemical reactions from US patents. https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed 2 May).
56. Kullback, S.; Leibler, R. A., On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951,** *22* (1), 79-86.
57. Li, M.; Zhang, Y.; Zhu, M.; Zhou, M., Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics: Sydney, Australia, 2006; pp 1025–1032.

Exploration of bioinformatic domain based on data mining, reaction predictions and enzyme promiscuity predictions

Electronic Supplementary Information

Chonghuan Zhang,[1] *Qianyue Zhang*[1] *and Alexei A. Lapkin*[1, 2 *]

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom

[2] Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd, 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore

∗ Corresponding author E-mail addresses: aal35@cam.ac.uk

Table of Contents

S1 Cofactors and free metabolites in KEGG database

The cofactors and free metabolites are adopted from Blaβ et al.[1] A full list of these molecules can be found in 'Cofactors and free metabolites' spreadsheet of the Excel file 'SI.xlsx'. These molecules were manually curated from Kyoto Encyclopaedia of Genes and Genomes database (KEGG) database. In total, it has 34 cofactors and 123 free metabolites (157 in total). The corresponding Reaxys molecule IDs are also listed in the table if these molecules can be found through Reaxys web interface.

S2 KEGG molecule similar compounds

Using SIMCOMP method,[2, 3] all similar compounds with a chemical structure similarity indicator Jaccard coefficient $JC \geq 0.85$ for KEGG molecules were recorded, and the full list can be found in 'Similar compounds of KEGG mols' spreadsheet of the Excel file 'SI.xlsx'. All molecules were given by their KEGG IDs and SMILES strings.

S3 KEGG One-to-one wiring reactions

The full list of 3,594 one-to-one wiring reactions with 3,079 recorded enzymes is shown in the 'One-to-one wiring reactions' spreadsheet of 'SI.xlsx'. All reactions are given with their reaction IDs/SMILES, reactants IDs/SMILES and products IDs/SMILES. All enzymes were given with their EC numbers, whilst the details of enzymes including their name, class names, pathways and *etc*. are shown in 'Enzyme details' spreadsheet of 'SI.xlsx'. In the columns of 'KEGG reaction', the enzymes catalysing more than one enzyme are identified as promiscuous enzymes in KEGG Enzyme database.

S4 Reaction SMILES tokenisation rules

The tokenisation of reaction SMILES follows the atom-wise rules to split reaction SMILES strings into tokens. The patterns used to split strings are discussed in Schwaller et al.,[4] and it is listed as follows:

"(\[[^\]]+]|Br?|Cl?|N|O|S|P|F|I|b|c|n|o|s|p|\(|\)|\.|=|#|-|\+|\\\\|\/|:|~|@|\?|>|*|\$|\%[0-9]{2}|[0-9])".

An example of tokenisation of reactions can be found in Figure 3a.

The one-to-one token-numerical index dictionaries of reaction SMILES, and enzyme amino acid sequences are shown in 'SMILES token dictionary' and 'AASeq token dictionary' spreadsheets of 'SI.xlsx'.

S5 First level of EC classification

Following the enzyme commission (EC) nomenclature rules[5], all enzymes can be specified by four levels, based on the reactions catalysed respectively. The first level of the EC number represents the major types of the reactions catalysed, and is categorised into seven classes (summarised in

Table **S1**).

Table S1 Summary of first level of EC classification.

Class	Name	Summary	Typical reaction
EC1	Oxido/ reductases	Catalyzation of oxidation/reduction reactions, transfer of electrons.	A+BH→AH+B, A→AO

EC2	Transferases	Transfer of functional groups (generally a glycosyl group or a methyl group) from a donor compound to an acceptor compound.	A+BC→AB+C
EC3	Hydrolases	Hydrolysis of C-O, C-N, C-C and some other bonds to form two products from one substrate.	AB→AH+BOH
EC4	Lyases	Non-hydrolytic cleavage of C-O, C-N, C-C and some other bonds. Addition or removal of functional groups from substrates.	RCOCOOH→RCOH, A-X+B-Y→X=Y+A-B
EC5	Isomerases	Geometric or structural changes (isomerization) within one molecule.	ABC→CBA
EC6	Ligases	Combine two molecules by hydrolysis of a diphosphate or triphosphate bond.	A+B+ATP→AB
EC7	Translocases	Assisting movements of molecules through membranes.	N/A

S6 Selected genes and amino acid sequences

The protein amino acid sequences of the enzymes are crawled from KEGG genes database. The selected genes and their respective amino acid sequences of the peptide chains translated from the genes are shown in 'Amino acid sequences' spreadsheet of 'SI.xlsx'. The lengths of the given peptide chains are also given in the last column of the spreadsheet. Statistics of the peptide chains lengths are show in Figure S1. The average length of amino acid sequences is 449, whilst

the median is 395. The length taken into the model input, 600, covers the length of up to 78% of the peptide chains, which is reasonable to retain most information.

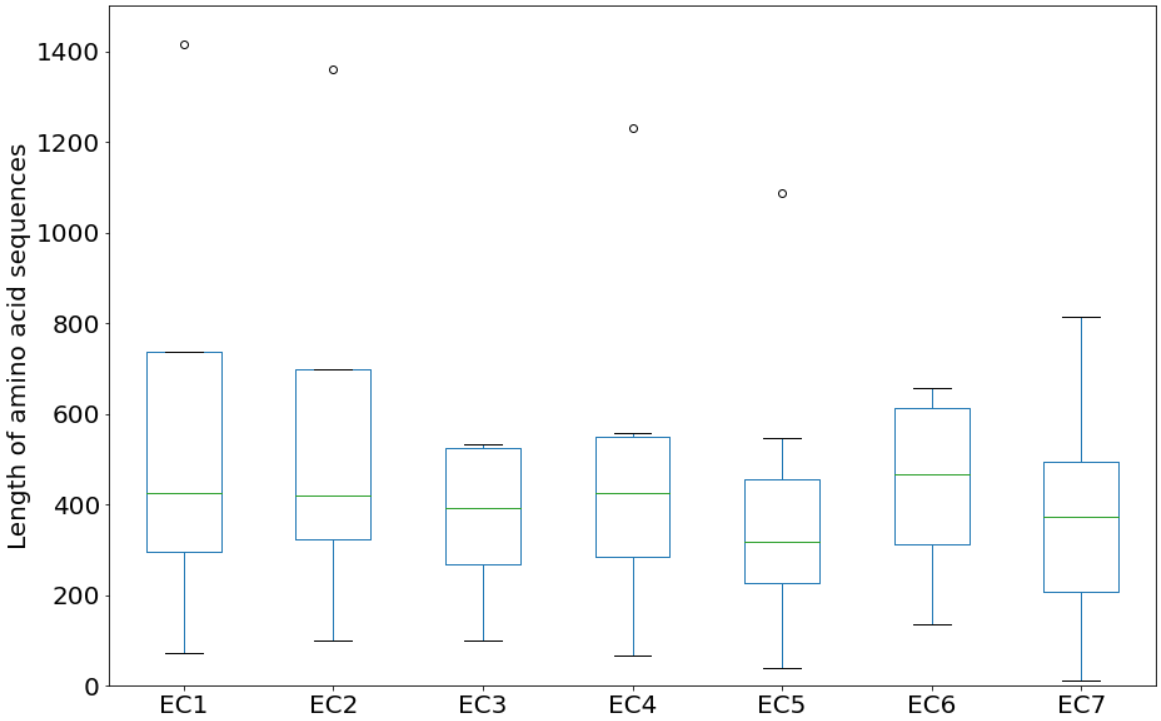


Figure S1 Statistics of amino acid lengths of the peptide chains, grouped by the seven first level of enzyme EC numbers.

The 21 amino acid types are differentiated from their side chain functional groups. The letter denotations of amino acid types in the amino acid sequences are shown as follows:

A: alanine, C: cysteine, D: aspartic acid, E: glutamic acid, F: phenylalanine, G: glycine, H: histidine, I: isoleucine, K: lysine, L: leucine, M: methionine, N: asparagine, P: proline, Q: glutamine, R: arginine, S: serine, T: threonine, V: valine, W: tryptophan, Y: tyrosine, X: unknown amino acid, O (or 0): vacancy.

S7 t-SNE visualisation of the biological reaction space

The t-SNE visualisation of the biological reactions by extended-connectivity fingerprints is shown in Figure S2. The visualisation resembles a ‘ball’ with points approximately equidistant

from its nearest neighbours, which indicates the failure of digitalisation of reactions into ECFP fingerprints.

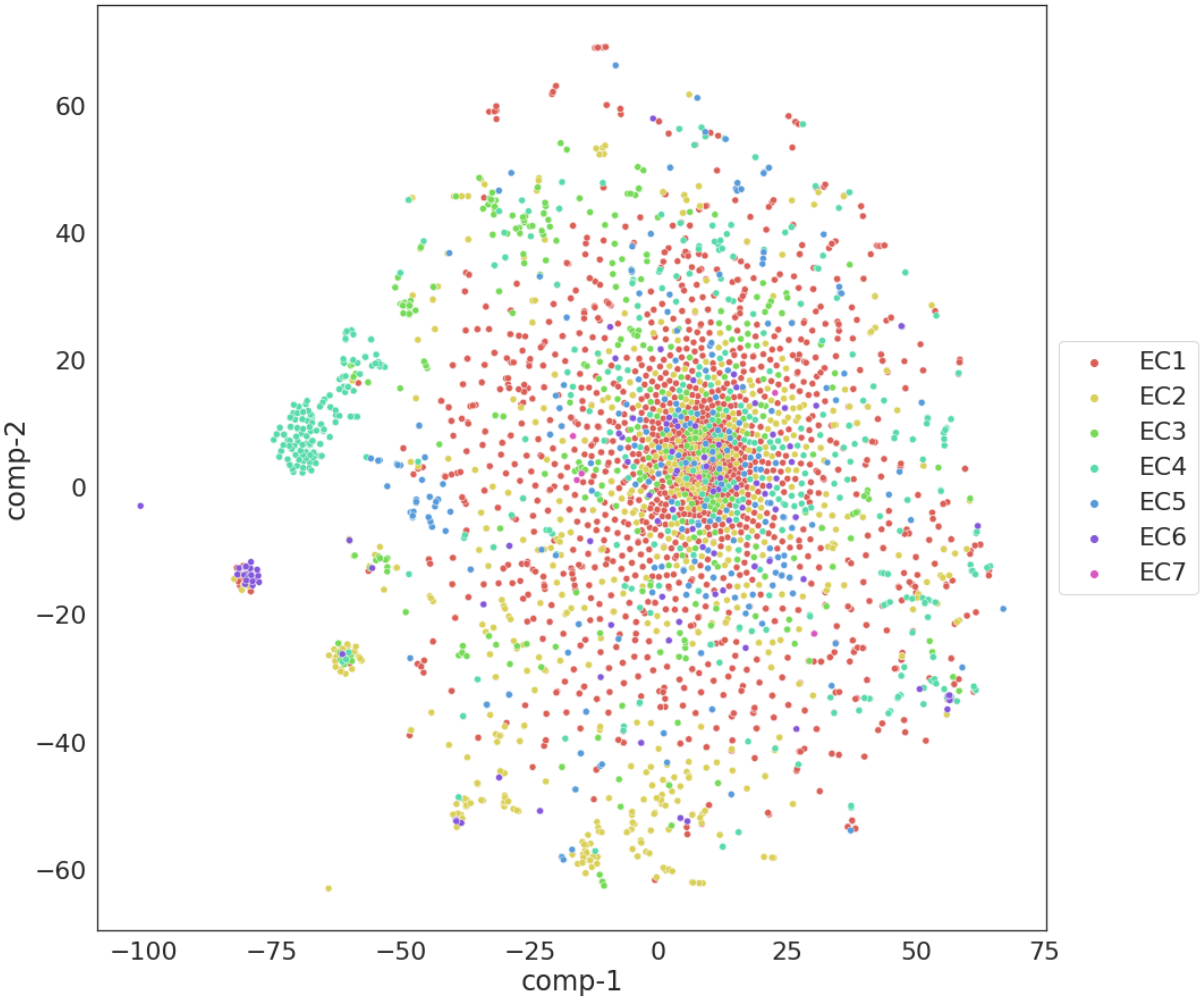


Figure S2 Two-dimensional t-SNE diagram to visualise biological reaction space from KEGG curated reactions using ECFP3 fingerprint.

S8 Transformer Model Result

The details of transformer model setting trials for the purpose of hyperparameter tunings, and their corresponding training and validation loss after 1000 epochs and test set KL divergence loss are shown in Table S2.

Table S2 Transformer model setting trials details

Trials	Dimension of embeddings	Dimension of feedforward units	Dropout	Number of hidden neurons for multihead attenton	Number of hidden layers	Batch size	Loss function	Training loss	Validation Loss
1	32	32	0	4	4	16	KL divergence	1.50	1.03
2	64	32	0	4	4	16	KL divergence	1.20	0.80
3	128	64	0	4	4	16	Cross entropy	1.12	0.48
4	128	64	0.1	4	4	16	Cross entropy	1.45	0.34
5	128	64	0.1	4	4	16	KL divergence	0.068	0.029
6	128	64	0.2	4	4	16	KL divergence	0.90	0.08
7	128	64	0	4	4	16	KL divergence	0.85	0.15

The transformer model predicted amino acid sequences from the test dataset are compared with the actual amino acid sequences of the enzymes in ‘model amino acid sequences’ spreadsheet of ‘SI.xlsx’, where Levenshtein distances between sequences are also given.

References

1. Blaß, L. K.; Weyler, C.; Heinzle, E., Network design and analysis for multi-enzyme biocatalysis. *BMC Bioinform.* **2017,** *18* (1), 366.
2. Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M., Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **2003,** *125* (39), 11853-65.
3. Hattori, M.; Tanaka, N.; Kanehisa, M.; Goto, S., SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.* **2010,** *38* (suppl_2), W652-W656.
4. Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T., "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018,** *9* (28), 6091-6098.
5. Enzyme nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Pp 862. Academic Press, San Diego. 1992 ISBN 0-12-227165-3. *Biochemical Education* **1993,** *21* (2), 102-102.