Concetta Press
Journal of Biomedical Informatics and AI

Glass Box Machine Learning to Aid the Design of Small Peptide Agonists from Very Sparse Data:

The Example of Erythropoietin Analogues

Barry Robson PhD DSc

Ingine Inc., Cleveland, Ohio, USA, and The Dirac Foundation, Oxfordshire, UK

Article Type	Original Research
Volume / Issue	1 / 1:4 (2026)
DOI	https://doi.org/10.5281/zenodo.20073063
Correspondence	Barry Robson, PhD DSc Email: barryrobson@ingine.com Ingine Inc., 11000 Cedar Ave Ste 100, Cleveland, Ohio
Received / First Decision / Accepted	April 19, 2026 / May 4, 2026 / May 7, 2026
Disclosure	Funding: None. Conflicts of Interest: None.

Abstract

This paper presents a perspective on interpretable approaches to peptide design when, as is often the case, initial data is very sparse. It provides a worked example with discussion for the very earliest stage that one may can consider a "hunch phase", when data is sparse and often circumstantial, and suspected activity and inactivity for many peptides is more correctly described as unknown. The very earliest steps involve bioinformatics and related aspects of computational chemistry approaches that are essentially standard, but their purpose is to generate all possible information for a final explanatory Glass Box AI approach involving use of a Theory of Expected Information developed for sparse data which is much less commonly described. They still have some novel or unusual features which make them well suited to that task, and glass box AI applied to peptides can itself be considered a recent form of bioinformatics. The present discussion can also be considered as an early step in the design of traditional therapeutics, i.e. small organic "in-a-pill" drugs. This is because biologically active peptides can provide clues for design of small organics, help establish laboratory assays, and provide important information as to the action of agonists and antagonists, and as to safety. Also providing early data are cases where peptides have reached the marketing stage, but still have disadvantages, even being withdrawn, so increasing the demand for more traditional therapeutics that may not have the same problems. In other cases, there is data from viable peptide therapeutics where use is confined to one or few countries. Erythropoietic peptides for treatment of anemia provide an example of all these cases.

Keywords: peptides; peptidomimetics; artificial intelligence; anemia; erythropoietin; sparse data

1 Introduction

1.1 Background

The design and development of peptides have long been recognized as providing guidelines that could significantly reduce early research and development costs of small "in-a-pill" organic compounds [1]. The COVID-19 pandemic

in January 2020 stimulated an interest in the applications of computer-based knowledge management and Artificial Intelligence (AI) to the rapid design of peptides as anti-viral agents when information was initially limited [2], and 2024-2025 saw a dramatic increase in the application of AI to therapeutic peptide design [3-6]. There, in many cases the initial data are relatively small as discussed below. Cases involving sparse data are widely recognized to be a difficult challenge for AI, but such cases are common across industries and there is currently great interest in developing methods to meet that challenge [7].

The traditional view has been that peptides as commercially viable pharmaceutical agents in themselves are limited by their sensitivity of the inter-residue peptide bond to gastric acidity and proteolysis, and with short plasma half-life by enzymic hydrolysis and renal clearance. Nonetheless, the US FDA has approved of the order of a hundred peptide-based drugs, not only of value in themselves but also contributing to the information available for the design of convenient small organic compounds to treat the same important diseases. A major advantage of considering peptides in early pharmaceutical research resides is their biological character, allowing the more direct application of bioinformatics. One may borrow useful information about molecular recognition, as contained in genomes, and from three-dimensional protein structures when available. The human body naturally produces over 7,000 known peptide types, refined by millions of years of evolution to achieve their very high degrees of molecular recognition and function. Also, they have just 20 naturally occurring units in a linear sequence which puts them in stark contrast to the hundreds of chemical groups and units typically comprising a medicinal chemists' toolkit, [8]. Starting with peptides also provides the opportunity to use various techniques in combinatorial peptide chemistry and phage display and mRNA display.

There are, however, limitations that can make it difficult for many researchers to obtain extensive data early. Combinatorial, phage, and mRNA techniques are expensive to set up, require expertise, and are more characteristic of "Big Pharma". For many smaller, e.g., academic groups, such techniques are not accessible, and results of their application in industry are not typically publicly available. Even for those researchers who do have access, early design of the experimental set-up still requires some kind of starting information, e.g., about the enzyme or receptor target. The approaches are also not perfect when used alone, e.g., due to unwanted effects of attached molecular labels to identify the compounds that bind to protein targets. Like the use of computational chemistry for small organic ligands, ligand-binding simulations can simulate the above experimental combinatorial approaches, but peptides are challenges because of their conformational flexibility [9].

1.2 Erythropoietic Peptides

Anemia affects an estimated 30% of the world's population, and some 10-11% in industrialized nations [10], but there are at the time of writing no erythropoietic drugs available in pill form for oral administration. Not all causes of anemia justify the use of erythropoietic agents. Some causes are easier to treat than others [10], and some that at first consideration might seem simple to treat, like bleeding peptic ulcer, have their complications when considering therapy [11]. The natural protein hormone erythropoietin (EPO) [10] was a blockbuster commercial product that helped launch the biotechnology revolution. However, it is expensive, requires continued injection [10], and inconvenient as it requires refrigeration, essentially prohibiting the patient from any extensive travel. EPO has some 450 million years of evolution, so it is a core player with many other actions, some of which may not always be wanted in clinical use. Success with small organic molecules to mimic EPOs [12] based on ideas similar to those used in the development of penicillin-like compounds [13] has been limited. For these reasons, there has been a significant effort to develop peptides, peptidomimetics, and other compounds more closely related to peptides with erythropoietic action [12,14-19]. Several peptides have been found to be erythropoietic and are mentioned in context below, but probably the most successful and almost certainly the earliest well-known peptidyl EPO receptor agonist was discovered by Wrighton and colleagues [17]. Using random phage display peptide libraries and affinity selective methods, they isolated small peptides that bind to and activate the EPO receptor. None of the above peptides clearly corresponded to sequences found in the primary sequence of EPO, though Cho and colleagues later found peptides with neuroprotective agonist effect by binding to EPO receptors that were drawn from the EPO sequence (see Discussion and Conclusions, Future Work). As was the case of GLP-1 and its semaglutide derivative, there has been the benefit of detailed structural data for erythropoietin and peptides deduced from it in complex with the receptor

to aid design. The experimental three-dimensional structure of EMP1 bound to the EPO receptor is of particular interest here [20,21], though the mode of binding is not the same as that of EPO.

Developing a commercial erythropoietic peptide (EPP) has been somewhat limited at the time of writing. It is the nature of drug discovery that this situation can change dramatically with new offerings prior to and after publication. The general principles discussed here will of course hold still, but there has arguably been a slowing down of efforts on EPP because past efforts have been subject to criticism [22-23]. Its use as a therapy is not always the best choice, because anemia can have several causes. There are always risks in new therapeutic compounds. Probably the most promising so far have been Pegmolesatide which is currently only marketed in mainland China with, at the time of writing no indications of being available in American or Europe, and Peginesatide, previously marketed in the U.S. and Europe but withdrawn due to safety issues. These and other erythropoietic peptides showing some promise have been peptidomimetics, meaning that there are chemical modifications to improve binding to protein target, reduce acidic and enzymic hydrolysis, reduce risk of immune response, reduce renal clearance, and improve pharmacokinetics such as ability to cross the blood-brain barrier. The emphasis has been on modifications known as PEGylation, cyclization, stapling (use of a covalent linkage between two amino acid sidechains, so forming a peptide macrocycle), and multimerization (e.g., Refs [15,17,18]). These are secondary steps in peptide development and are not considered here, except for a brief discussion of peptides made largely or entirely of D-amino acids. These are peptides that can map structure and function more closely to L-amino acids while having more desirable pharmacokinetic properties and can be readily applied to large peptides and even proteins [24].

1.3 Special Considerations

The scenario exemplified by EPPs requires some special considerations discussed primarily in Theory and Methods Section 2. An important one is that great care must be taken in introducing peptides assumed to be *inactive* as EPPs. In the past, researchers applying analysis and prediction techniques to small organic molecules as potential drugs boosted the amount of input data by including essentially arbitrary compounds that were *presumed* to be inactive, i.e. highly unlikely to be active by chance but were really unknowns [8]. More recently, standard databases have been developed for that have not only been proven inactive but are also closer to those which are active [8]. This is important to define crisply the classification boundary between activity and inactivity of any chemical substance, but intuitively it should be particularly important for peptides with 20 different chemical groups falling into some 5 classes of properties (e.g., large hydrophobic sidechain). Changes in activity with structure such as a single amino acid class change is likely to be more sudden, as indicated by effects of single amino acid mutations [25]. At the same time, we are interested in the case when true inactivity data is also sparse. This suggests that one picks inactive peptides that are not arbitrarily chosen but *related to known active compounds by evolutionary or research processes*. Estimates from many early computational chemistry and screening methods (e.g., Ref. [13]) predict active compounds that turn out to be inactive in 90-99.9% of cases make the hunch that they are inactive reasonable.

2 Theory and Methods

2.1 Overview

The core features in this worked example involve the use of "Glass Box" algorithms that may be considered as Large Probability Models. These involve some lesser-known mathematical aspects and a reviewer suggested that a simplified overview might be provided in a "Biological Intuition" section (Section 2.7). These models replace arbitrary learned weights of neural net methods with large numbers of annotated joint (multifactor) probabilities and also provide simplified explanatory models [8,26]. See Section 2.5. Here partially arbitrarily distributed weights of neural net methods are replaced by high-dimensional probabilities or odds [26]. Applications to molecular design including DiracSmash [26] have more recently included several studies including a rapid response to the rise of Covid-19 [27-30] and for small organic compounds alongside contemporary AI approaches. This has also been to provide explanatory models and to guide synthesis from a typical set of initial studies by the medicinal chemist [8]. There, sparsity was to some extent compensated for by the entries that were mostly compositional, i.e. relating to the type and number of a large variety of chemical groups rather than by any notion of position in the formula. In

the present study, the theory and method are such that data used is not presented in a compositional way but depends on the type of amino acid residue at each location in aligned sequences, as is appropriate to peptides.

2.2 Hunch Strategy and The Theory of Expected Information

Purely irrelevant, arbitrary, or random sequences presumed to be inactive will tend to be ignored by the Theory of Expected Information (TEI) used here (e.g., refs [8,26]) because it weights contributions in a natural way according to the amount of data available. For example, the impact of 1 out of 5 observations would be included but carries a lot less expected information than 100 out of 500. Also, weak homology helps build the alignment and inclusion of irrelevant or arbitrarily chosen sequences presumed inactive introduces many insertions (gaps) to obtain alignment, so it becomes even more obvious that including irrelevant or random peptide sequences assumed inactive is not a useful strategy here because specific residue types occurring at aligned positions will be seen even fewer times. Of course, data on inactivity is important where available (e.g., Ref [19]).

In Section 1.3 it was noted that drug discovery has made frequent use of data on compounds that are assumed inactive but are more correctly unknowns. This is inevitable in the present "hunch" model but the matters should be seen as representing some judicious Bayesian prior belief by the researcher. The TEI addresses not only sparse data but also inclusion of prior belief or knowledge from other sources. For example, conditional information, appropriate to predictions and estimation of the Likelihood Ratio etc. as discussed later below, is as follows.

$$I(a|b) = \zeta(s=1, n[a,b]+v[a,b]) - \zeta(s=1, n[b]+v[b]) \quad (1)$$

Here a and b can each be joint events and in principle each a logical expression, describing a situation that is observable and countable. Parameter n relates to observed frequencies, i.e., counts of events in the dataset used such as [a,b] and [b], and v relate to virtual frequencies computable from prior Bayesian belief or external knowledge. The zeta function emerges naturally from integration of information measures over Bayesian posterior probabilities (see Ref [8] for discussion and references) and is defined as follows.

$$\zeta(s=1, (n+v)) = 1 + ½ + ⅓ + \ldots. + 1/(n+v) \quad (2)$$

A simple example of a prior degree of belief re-expressed as a virtual frequency is to compute it from the Bayesian degree of belief as a probability times the total amount of data N. A good estimate is given by the following when mutual information (the natural log of an association constant) is required. Here n corresponds to the observed number of events n[a,b] and e[a,b] = n[a]n[b]/N is the expected number, in the familiar chi-squared sense, i.e.

$$I(a; b) = \zeta(s=1, n[a,b]) - \zeta(s=1, e[a,b]]) \quad (3)$$

The above is a particular basic knowledge reference point to which further virtual frequencies may be added to n[a,b] and e[a,b], but it is sufficient that the values of (n+v) for Eqn. 1 and (n+v) and (e+v) for Eqn. 3 are driven by the appearance of records in the database. To provide input to obtain the analogue of a training set which takes account of v as further external knowledge, a variety of knowledge gathering tools, bioinformatics, and three-dimensional structure data concerning interactions (here between the target protein and erythropoietin and peptides found erythropoietic), are important [33-38]. In this case, the sequences represent degrees of belief about factors dictating activity or inactivity that are included in the alignment and appear as the effective count (n+v) or (e+v).

An unusual feature is the idea that sequences possibly related by evolution and serving some other function by some criteria but likely to be inactive as therapeutic agents are used. Studies have suggested that receptors often have weakly homologous sequences to their peptide or protein ligands and this has been put on a rational basis by Dwyer [31]. Peptides so defined could in principle be treated as potentially active peptides but extensive substation into different residue classes have occurred in the evolutionary process [25]. The assumption of inactivity is further boosted when Dwyer's principle is combined with information from structural data (Section 2.3). Evolved activity relevant to therapeutic effect would appear unlikely as it involves complementary residue-residue interactions, not residue matches. However, general hydrophobic-hydrophobic, hydroxyl-hydroxyl interactions, and cation bridges between sidechains of same charge could conserve binding features relevant to activity [25].

2.3 Sidechain-Sidechain Interactions in Peptide Binding

Particularly persuasive assumptions of activity or inactivity can be derived from experimental three-dimensional structure data providing insight into the physical nature of relevant ligand-receptor interactions. While such analyses are usually considered as implying the use of computational chemistry [33-38], calculation of free energy with entropy may be less reliable than empirical measures of binding between ligand and receptor residues [30,36,37]. A detailed molecular dynamics simulation would add refinement to the results of such a study [38] but requires very substantial resources to compute reliable entropies of the protein-ligand-solvent system, unsuited to an initial feasibility study. The method used here [30] appears to be particularly effective for present purposes. The sidechains of peptide-receptor interactions are essential for molecular recognition because they confer specificity. The method treats the sidechain more precisely as the outer part of the sidechain which carries molecular recognition, from the C_γ carbon outward except for glycine and alanine that use C_α and C_β atoms respectively. In the examples in Results Section 3.1, the digits under the amino acid residues in standard one letter (IUPAC) code are the exposure scores, ranging from 0 to 9. X means a degree of exposure score of 10 or more, and again in practice almost always means 10, especially in the present study. A useful procedure [30] is to explore the exposure scores for the receptor in the absence of the coordinates for the ligand, and then when the ligand coordinates are replaced. Even more importantly, one can explore the exposure scores in the absence of the coordinates for the receptor, and then when the receptor coordinates are replaced. Where residue sidechains are buried by the ligand-receptor interaction, the exposure score is expected to fall, and a fall of the score by 3 is usually considered particularly significant.

2.4 Source Peptide Sequence Data

While the maximum datasets studied comprised 5000 sequences assumed inactive from aligned weakly homologous sequences or irrelevant or random sequences, playing at best a peripheral role in the sense discussed in Section 2.2. To provide the reliable small core portion of the data, Refs [12-15,17-20,32,39-45] were important sources, or provide references to real experimental activity data used in this study (they also provide information useful to help for the design of analogues with D-amino acid residues [24,46]). The curated and aligned dataset for data accessible at the time of the study and used for training by the Glass Box approach comprised 111 sequences. This was considered appropriate to be a good test of very sparse data analysis. The method of obtaining appropriate references for that data was automatic searching of the Internet somewhat in the manner of Large Language Models but including links with date and time stamps to primary and other sources [27-30]. The first-pass agonists of Wrighton et al. [17] were important. They are represented by a 14-amino acid disulfide-bonded, cyclic peptide with the minimum consensus sequence YXCXXGPXTWXCXP, where X represented positions allowing occupation by several amino acids [17]. Of particular interest is a peptide called EMP1, now considered as TYSCHFGPLTWVCKPQ [18] with diglycine GG added at the N- and C-termini, making it a 20-residue peptide overall. Methods based on partially summated zeta functions for making use of sparse data based on the TEI [47,48], can modify the counts based on the sequence data in a natural, theoretically sound way to include prior belief but that was not done here for active peptides. Other data, notably such as the empirical parameters representing residue exposure and free energy estimates of residue interactions are beyond present scope were obtained from Refs [36,37,49].

2.5 Knowledge Element Tags, Queries, and Query Guidance

Several “Glass Box” Machine Learning methods are available in the suite developed by the author, but those used in this study were QFANO, DiracMiner, and importantly DiracSmash [26]. All provide association (disproportionality) constants and a fuller set of relevant measures, such as odds, respectively. A common feature is that that they generate probabilistic knowledge elements or “tags” for patterns found one or more times. These are the annotated probabilities mentioned in Section 2.1. They follow the Dirac notation and algebra in the sense described in Ref [26]. Many thousands of these can then be assembled into a prediction network. An example tag generated by DiracSmash in one study is as follows.

< 'Active':='eq yes' Pfwd:=1.0000 Ofwd:=2.0368 Efwd:=0.2500 | **if**:=(assoc:=7.3703, count:=1, factors:=(3,5)) | '12':='eqP' '17':='V' **with** '25':='L' '10':='W' Pbwd:=0.2883 Obwd:=13.2823 Ebwd:=1.8750 >

These tags contain important additional information for assembling prediction networks and making and verifying predictions, and for analysis and reusability, explanation of which is beyond present scope, but note that Obwd (odds backward) corresponds to the Likelihood Ratio discussed in Section 2.6. Also, there are *attributes* 'Active':='eq yes' followed by '12':='eqP' '17':='V' with '25':='L' '10':='W'. These represent 'Active':='eq yes' as the target to predict as active when valine occurs at position 17 '17':='V', leucine at 25 '25':='L', and tryptophan at position 10 '10':='W' (note the use of the standard IUPAC one letter code for amino acid residues).

QFano and DiracMiner are unsupervised datamining methods: they perform a high dimensional (multifactor) data analysis of any structured dataset to generate the above measures. Particularly important for predictions, however, is DiracSmash that produces a prediction from a query when given as data a named file with a specified range of records as the "training" or test set as appropriate. The prediction target in the present case is active:= 'eq yes' versus active:='ne yes' meaning active:=no, followed by specification of several attributes '12':='eqP' '17':='V' **with** '25':='L' '10':='W'. The attributes preceding the 'with' operator are required by the query to be present in all tags, and those following that operator are generated by unsupervised data mining (with less sparse data they may extend up to 4 , 5, or 6 further attributes dependent on version). An important qualification of the above is that the query can specify independent attributes of which at least one must appear to the right of 'with'. By "independent" is meant that they are to be treated as interdependent with the target and previous attributes, but treated independently of each other. Even when data is sparse, a query can contain many attributes. For example, a major query of interest in the study, discussed in Results Section 3, is (in briefer notation) active:=yes, 3:=G, 4:=G, 5:=T, 6:=Y 8:=C, 9:=H, 9:=K, 12:=P, 14:=T, 15:=W, 15:=F, 15:=W, 16:=H,17:=V, 18:=C, 20:=P, 20:=A, 21:=V, 21:=G, 22:=G, 23:=G, 25:=L, 26:=R, 27:=S, 29:=x, 34:=A. Here it was with '12':='P', and '17':='V' as entered as interdependent attributes and the rest as independent attributes. If the data will not support such a query, the user is notified. Overall, the DiracSmash algorithm proceeds in the following steps.

1. Read the query as the target (e.g., active:=yes), plus a list of descriptors (attributes) such as 3:=G which are interdependent with each other and the target, plus a list of similar descriptors that are independent of each other but interdependent with the target and the interdependent descriptors.

2. Perform a preliminary fast mining to obtain self-probabilities such as P(3:=G) and various values required for speeding algorithms and memory management (of the implied combinatorial explosion that generates large numbers of tags) [26].

3. Perform an exact traditional calculation of Predictive Odds (PO) and Likelihood Ratio (LR) for the target (active versus inactive) based on direct counting of activity and inactivity with the descriptors on the list of interdependent variables. If there is insufficient data for the exact calculation, so requiring use of independence assumptions, weaker descriptors on the interdependent list are determined and moved to the independent list.

4. Generate a reduced data set (fewer records) in which all records include all the descriptors on the list of interdependent variables.

5. By mining the reduced data set (from step 4), build an inference net [26] composed of all generated tags (see above) to predict the PO and LR using information from preliminary mining (step 2) so that they are equivalent to the result of analysis of the original full dataset. A typical inference net constructed in the present study typically contained 126,925 tags but occasionally as few as 4,323 tags.

6. Compare the PO and LR predicted by the inference net with that obtained by exact traditional calculation (step 3) and calculate correction factors for applying the "perturbation method" [26]. This is because any inference method (and implicitly current AI methods) is an estimate: it makes many independence assumptions that neglect interactions.

7. Apply the descriptors in the independent list by removing tags that do not contain at least one descriptor on the independent list. A typical inference net in the present study at this stage contained 27,657 tags but minimally 942 tags.

8. Apply the correction factors from step 6, regularize and normalize the inference net and show that the probabilities involved are "coherent", primarily meaning that they satisfy Bayes' Rule, here P(active | descriptors) P(descriptors) = P(descriptors | active) P(active). Report PO and LR.

9. Generate a simplified explanatory model and test it on a user-selected test set. There are several tunable options of model right up to the full calculation used in step 8. The simplest option used in the present study and elsewhere (e.g., Ref [8]) is that a record predicted as active contains a minimum fraction or more of the descriptors in the query, interdependent and independent descriptors being given equal weight. The minimum fraction is the threshold underlying ROC curve construction [26].

For fuller details, see Ref [26]. Although the algorithm is general, it was originally developed for the case when the query was the description of a specific patient, on which a prediction would be made using information mined from many patient records. For the present study, the appropriate query (in terms of separate interdependent and independent descriptors) is not always so obvious. To help select entries for the query, an initial phase of unsupervised datamining, i.e., without any query, is performed to provide the query as top-ranked associated attributes. This can be provided by DiracSmash itself, though typically it is done by a separate code lacking the role of the query, for efficiency, and which of obtaining high-dimensional information, probabilities etc. involving many attributes. The mutual information here is based on Eqn. 3. The order of ranking of joint events for s=1 is preserved by using other values of s in $\zeta(s, n)$ which express various surprise measures, but Eqn. (2) uses s=1 in this study and is taken as the *definition* of information, related to information as logarithms of probabilities. See Refs [47,48] for discussion of the Euler-Mascheroni "constant" Euler-Mascheroni constant $\gamma = 0.57721\ldots$, such that $\log_e(n) + \gamma = \zeta(s=1, n)$, *Lt* $n \rightarrow \infty$. It is more correctly described as a function, but in either case it essentially cancels in Eqn. 1 and Eqn. 3. In the present study of very sparse data many investigations involved just one or very few observations, so the probabilities and odds and association constant on the tags used the option of being reported and used based on the Theory of Expected Information, i.e., functions $\zeta(\ )$ as described above. Note that this aspect is extremely important for the use of very sparse data. Replacing the functions $\zeta(\ )$ by traditional measures based on logs of ratios of counts in preliminary studies showed deterioration of predictions in the range of 75% to 92% accuracy to 55% to 70%.

2.6 Measurement of Predictive Capability

There are two kinds of predictions made by DiracSmash. Predictive capability is tested both for the "Large Probability Model" that imply many tags, and for simplified models. The simplified models are useful because, combinatorically, many thousands or millions of tags (but only up to about 126,925 in the present sparse data study) are generated from a test set. Tags are individually intuitively comprehensible to the human researchers but collectively overwhelming. The optional explanatory model used was the simplest of the simplified models. It involves DiracSmash calculating what fraction (e.g., 70%) of the descriptors such as 9:=H, 9:=K, … in the query are actually found in each specific peptide. It is then compared with a minimal required fraction to predict Active:=yes, otherwise Active:=no. For the "training" set and for the test set overall, that minimal fraction used represents an adjustable decision threshold, so having established the numbers of true and false positives and true and false negatives for predicting activity, a standard ROC curve plotting true against false positives, as a result of increasing the thresholds at small intervals from 0% to 100%, can be constructed [8,26]. Note that the use of the method is objective compared to much use of the ROC curve in the literature (see Ref [26] for discussion). The optimal threshold is established only for the "training" set and then becomes a fixed part of the simplified predictive model, which is then applied to the *test set* for independent validation which is now uninfluenced by any prior knowledge of the test set. The fixed threshold is established from the "training" set by finding the optimal value of the following for predictions. This can easily be done by an exhaustive search as it essentially corresponds to tracing out the adjusted decision threshold that is responsible for plotting out an ROC curve [26].

$$\text{Quality\%} = (\text{accuracy\%}) - \mathbf{w}_{\text{balance}} \mathbf{X} (\text{sensitivity\%} - \text{specificity\%})^2 \quad (4)$$

Multiplier $\mathbf{w}_{balance}$ is set at 0.01 in the present studies. Quality% favors solutions in which sensitivity and specificity are reasonably balanced, which is the usual preferred case unless there are specific reasons for emphasizing sensitivity over specificity or *vice versa*.

Other well-known measures are positive Likelihood Ratio LR+ and negative Likelihood Ratio LR- which in the present case would correspond to predicting active and inactive respectively. LR+ = LR, the basic LR as discussed above. While one may use LR+ or LR- to optimize performance in a ROC curve study, the use of Diagnostic Odds Ratio DOR = LR+/LR- is currently popular. Expressed in terms of the incompletely summated zeta function it is as follows.

$$DOR \approx e^{\zeta(s=1,TP) + \zeta(s=1,TN) - \zeta(s=1,FP) - \zeta(s=1,FN)} \quad (5)$$

Another estimate of predictive capability can readily be given by the LR, and it has special relationship with the use of Equations such as Eqn. 4 and the point on the ROC curve at which accuracy = sensitivity = specificity. Eqn. 4 estimates this with some "slack". Recall first the following.

$$\text{accuracy} = (\text{sensitivity})(\text{prevalence}) + (\text{specificity})(1 - \text{prevalence}) \quad (6)$$

When sensitivity = specificity, the above implies the following.

$$LR+ = \text{sensitivity} / (1 - \text{specificity}) \rightarrow \text{accuracy} / (1- \text{accuracy}) \quad (7)$$

$$LR- = (1 - \text{sensitivity}) / \text{specificity} \rightarrow (1- \text{accuracy}) / \text{accuracy} \quad (8)$$

At the balance point of sensitivity = specificity then for LR+ (i.e., LR) the following applies.

$$\text{Pred} = LR / (1+LR) = \text{accuracy} = \text{sensitivity} = \text{specificity} \quad (9)$$

In typical use, the DOR and Pred = LR(1+LR) for LR > 1 usually vary rather little along the ROC curve in the approximate vicinity of the point of inflection. The DOR typically increases only approximately 6%-7% at thresholds of 5% and 95% (i.e. near the two ends of the ROC curve) in other studies.

2.7 Biological Intuition

Previous papers describing the method (e.g., Refs [8,25]) and the formal basis of the tag in Section 2.5 have their roots in the notation and algebra of Dirac for quantum mechanics, but for the present paper those aspects of the technology can be ignored when consulting the source references cited. That is because the inference nets constructed are odds inference nets of relatively simple form considered as acting in only one "direction", the direction concerned with Likelihood ratio LR.

What cannot be ignored is the use of zeta functions $\zeta(\ )$ because these are required for the management of sparse data, including combining it with more plentiful data. Eqns. 1-3 and associated discussion are given here because this paper essentially seeks to suggest a formal approach to sparse data in pharmaceutical research. Recall that, ultimately, the general problem addressed is that seeing 2 out of 5 occurrences of something, i.e., 2 that it is the case and 3 that it is not, provides less information than 200 out of 500 observations. Considering a test procedure as "underpowered" and basically stopping and awaiting more data is the standard procedure in the kind of frequentist statistics taught in high school, but that is not an option in the present study. The situation is somewhat like the situation in a Court of Law when a great deal of weak or circumstantial evidence can add up to outweigh the judgement that would be made without it. The amount of information that probabilities and odds carry when the data used to calculate them were sparse are not arbitrary empirical and guesses at numbers but formally established quantities as the expected value of log ratios of probabilities based on the use of the well-known Bayes Rule. The full theory with $s > 1$ in $\zeta(s=1,\ldots)$ is important for future developments of the present project where there is need for other kinds of surprise measure and for its combination with Dirac's quantum theory adapted as a data analytic tool, but for the present paper it suffices to say the following. If a pattern in a peptide sequence is seen as associated with activity for the first time then the information obtained is 1 nat (natural unit), the second time we gain 1/2 nats as it is less of surprise, the third time add 1/3 nats, and so on till we complete n observations and overall gain 1 + 1/2

+ 1/3 +…+ 1/n nats. Then we may do that again for inactivity, say obtaining 1 + 1/2 + 1/3 +…+ 1/m nats and subtract that sum from the first. This calculation quickly converges to natural log(n/m) as data, i.e., as the number of observations n and m, increases, but use of that traditional measure is not desirable. As discussed in Section 2.5, preliminary studies showed deterioration of predictions in the range of 75% to 92% accuracy to 55% to 70%. This is because many contributions involve very small counts and would disproportionately contribute to the overall information as a weight of evidence. The added v in $\zeta(s=1, n+v)$ (Eqn.1) is a formal means of introducing belief and prior or external knowledge. It is equivalent to including active and inactive sequences as records in the analyzed data, where patterns in them contribute the counts v.

3 Results

3.1 Preliminary Studies

Preliminary test studies were performed on an aligned set of 27 erythropoietic peptides related to those of Wrighton and colleagues [17,18] of which TYSCHFGPLTWVCKPQ (with diglycine GG added at the N- and C-termini, making it a 20-residue peptide overall) is a representative member. They defined a disulfide-bonded, cyclic peptide with the minimum consensus sequence YXCXXGPXTWXCXP, where X represented positions allowing occupation by several amino acids [17]. Table 1 was obtained for preliminary associations expressed as mutual information measures, up to associations of 3 amino acid residues at a time. Though only the top 20 associations are shown, attributes mentioned in associations including Active:=yes of 2.56 nats (natural logarithmic units) or more are selected as the DiracSmash query. That means a threshold of an association constant of 13 or more, i.e. the number of occurrences of the cluster compared with what would be expected on a chance basis (based on individual probabilities of the attributes). Note that the numbering of positions in the method, e.g., 20:=P, relies on using

information	event1	event2	event3	event4
2.15	20:=P	3:=G	4:=G	Active:=yes
2.11	18:=C	20:=P	4:=G	Active:=yes
2.1	18:=C	3:=G	4:=G	Active:=yes
2.07	3:=G	4:=G	8:=C	Active:=yes
2.06	20:=P	4:=G	8:=C	Active:=yes
2.05	18:=C	20:=P	3:=G	Active:=yes
2.02	17:=V	3:=G	4:=G	Active:=yes
2.01	20:=P	3:=G	8:=C	Active:=yes
2.01	18:=C	20:=P	8:=C	Active:=yes
1.99	3:=G	4:=G	9:=H	Active:=yes
1.96	17:=V	20:=P	4:=G	Active:=yes
1.94	17:=V	20:=P	3:=G	Active:=yes
1.94	16:=x	3:=G	4:=G	Active:=yes
1.94	14:=T	3:=G	4:=G	Active:=yes
1.93	15:=W	18:=C	5:=T	Active:=yes
1.92	15:=W	18:=C	20:=P	Active:=yes
1.91	18:=C	4:=G	8:=C	Active:=yes
1.91	20:=P	4:=G	9:=H	Active:=yes
1.91	18:=C	5:=T	8:=C	Active:=yes
1.91	15:=W	20:=P	4:=G	Active:=yes

Table 1. Mutual information in natural logarithmic units for patterns found in the initial small collection of peptides of interest for developing erythropoietic peptides.

sequences aligned in the same consistent way rather than the formula representing the actual structure of the peptide with deletions removed. It is thus important to allow a sufficient but not excessive number of deletions. Deletions do not appear in the top 20 (Table 1) except for 16:=x in the 13th row of Table 1. In practice, the algorithms used represented deletions in bioinformatics sequences by the lower-case character 'x', which is not to be confused with 'X' usually used in bioinformatics to indicate that many different amino acid residues will do at that locus. Because of the original clinical applications, X' was managed as if it were an unknown value in the data, i.e. it was presented in the data as a blank or by the word 'unknown'.

Extension from 27 to 111 records was done (a) by using BLASTP to find homologous but not identical sections of protein sequence 36 residues long including deletions or shortening to 16 residues by end residue deletions, (b) making at least one manual radical non-conservative substitution, e.g., a hydrophobic leucine L to charged aspartic acid D or vice versa, and (c) by applying the following techniques (Section 3.2 to 3.5).

3.2 Extending the Data Set: A. Sidechain Interactions Important in EMP1-Receptor Binding

Attention was given to human EPO, the human receptor, and EMP1, using Protein Data Bank Entries such as 1EBP, complex between the extracellular domain of erythropoietin (EPO) receptor [ebp] and an agonist peptide [EMP1], see Fig 1, relating to Ref [25]. See also PDB entries for 1EER, the crystal structure of human erythropoietin complexed to the receptor at 1.9 Angstroms, and 1ERN, native structure of the extracellular domain of erythropoietin receptor at 2.4 Angstroms. The following code names areas used in those entries. The primary sequences of particular interest are as follows.

>1EER_1|Chain A|ERYTHROPOIETIN|Homo sapiens (9606)

APPRLICDSRVLERYLLEAKEAEKITTGCAEHCSLNEKITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRGQALLV
KSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISNSDAASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGD
R

>1EER_2|Chains B, C|ERYTHROPOIETIN RECEPTOR|Homo sapiens (9606)
REFPPPNPDPKFESKAALLAARGPEELLCFTERLEDLVCFWEEAASAGVGPGQYSFSYQLEDEPWKLCRLHQAPTARGAVRF
WCSLPTADTSSFVPLELRVTAASGAPRYHRVIHINEVVLLDAPVGLVARLADESGHVVLRWLPPPETPMTSHIRYEVDVSAGQ
GAGSVQRVEILEGRTECVLSNLRGRTRYTFAVRARMAEPSFGGFWSEWSEPVSLLTPSDLDP

>EMP1
GGTYSCHFGPLTWVCKPQGG

It is helpful that the X-ray structure of the EPO receptor complex with EMP1 (PDB entry 1EBP) showed an approximately overall symmetric EPO receptor homodimer complex with a EMP1 homodimer as ligand. See Fig. 1.

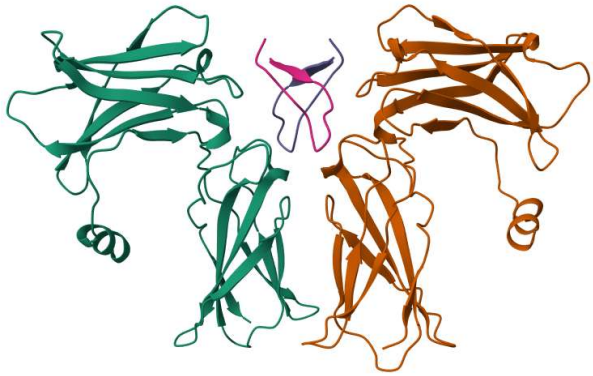

Figure 1. X-Ray Crystallographic Structure 1EBP of the EPO Homodimer Receptor Binding the Synthetic Agonist EMP1 Homodimer, showing the high degree of symmetry.

For analysis of interactions for drug development it is almost always sufficient to study interactions between one member of the receptor homodimer and one member of the EMP1 ligand homodimer. The terminal GGs were not shown and so presumably disordered (conformationally highly flexible) [30]. Although it is reasonably presumed that this represented an active agonistic form of the complex, the mode of binding and the conformation of the

receptor were considerably different to the asymmetric structure obtained with EPO as ligand. Middleton et al. [19-21] correspondingly noted that there were some determinants of the of the EPO receptor that are important for binding EPO and EMP1, but also some unique to binding EMP and EMP1.

Exploring interactions involved in ligand-receptor binding is primarily to be seen as a help in establishing an initial alignment frame. The first of the following two sets of output show the exposure score for the sidechains of chain A of the receptor in the homodimer (chains A and B) without EMP1 homodimer (chains C and D) coordinates. The second shows the sidechain of chain A exposures with the EMP1 homodimer replaced, i.e. the full set of reported coordinates for the PDB entry 1EBP. The results representing this process for chain B of the receptor are omitted for brevity since they are similar as the symmetry of Fig. 1 would suggest, especially in the region of EMP1 homodimer binding. The departure from symmetry is about the same as the EMP1 agonist as will be discussed next, which is the more important molecule for present purposes, i.e., as far as design of novel peptides and peptidomimetics is concerned. An additional feature is a second row of exposure scores called "smoothed": each score at a locus I in the residue sequence is the average from i-5 to i+5 and is useful in indicated an overall exposed surface loop. The string NYS in the amino acid residue sequence represents a covalent glycosylation site but no such site was involved in the binding of EMP1. Those expected to have a significant reduction on binding, by visual examination of the graphics display of the ligand-receptor complex, are shown in italic and underlined. They represent main parts of the 5 visibly rather obvious EMP1 binding loops that several authors have discussed and which were apparent in Fig. 1. Those sidechains appearing to have particularly strong interactions on the above basis are shown in **bold font**.

```
KFESKAALLAARGPEELLCFTE<u>*RLED*</u>LVCFWEEAASAGVGPGnYsFSYQL 50 (59) +59
XXX69555989X9X7927350554833334154878XX666863632233
XX9988888888887666555444444455566667777655445445
<u>*EDEP*</u>WKLCRLHQAPTARGAVRFWCSLPTADT<u>*SS**F**VPLEL*</u>RVTAASGAPRY 100 (109) +50
7938468491555938784350526368852748573425132587795 3
5556666565666656555555555555566665545444445555666 6
HRVIHINEVVLLDAPVGLVARLADESGHVVLRWLPPPE<u>*TP**MT**SH*</u>IRYEVD 150 (159) +50
7754521721536515727386399766332617615X369677151323
6554444443444445555566666555554454555666655454444 4
VSAGNGAGSVQRVEILEGRTECVLSNLRGRTRYTFAVRARMAE<u>*PS**F***</u>GGFW 200 (209) +50
3655XX77XX765X37437465646X587X48343213053379742364
5556777788777766555565666676666554433344444555555 5
SAWSEPVSLLT 211 (220)
2752X745574
55555565665

KFESKAALLAARGPEELLCFTE<u>*RLED*</u>LVCFWEEAASAGVGPGnYsFSYQL 50 (59) +59
XXX69555989X9X7927350553833334154878XX666863632233
XX9988888888887666555444444455566667777655445445
<u>*EDEP*</u>WKLCRLHQAPTARGAVRFWCSLPTADT<u>*SS**F**VPLEL*</u>RVTAASGAPRY 100 (109) +50
7938468491555938784350526368852524473425132587795 3
4556666565666656555555555555555554444344445555666 6
HRVIHINEVVLLDAPVGLVARLADESGHVVLRWLPPPE<u>*TP**MT**SH*</u>IRYEVD 150 (159) +50
7754521721536515727386399766332617615X341345151323
6554444443444445555566666555554454444444444333334
VSAGNGAGSVQRVEILEGRTECVLSNLRGRTRYTFAVRARMAE<u>*PS**F***</u>GGFW 200 (209) +50
3655XX77XX765X37437465646X587X48343213053378632364
5556777788777766555565666676666554433344444545545
SAWSEPVSLLT 211 (220)
2752X745574
55555565665
```

Considering the EMP1 homodimer, i.e. the complete coordinates for the whole ligand-receptor complex (the second set of scores) only has a strong effect on decreasing exposure in two cases, *SS**F**VPLEL* for which exposure scores 74857342 fall to 52447342 in which only the phenylalanine (F) sidechain shows a fall from 8 to 4. The second, *TP**MT**S*, shows a fall from 369677 to 341345 where the fall for the methionine (M) sidechain from 9 to 1 is particularly marked. Experimental substitution studies by Middleton and colleagues [19-21] indicate that Phe (F) 93 and Phe (F) 205 are important for both EPO and EMP1 binding, though in the present study *PS**F* the exposure scores fell only marginally, from 974 to 632, i.e. only by 2 for that Phe 205. Met 150 is not important for EPO binding but is important for EMP1 binding, supporting the marked fall for that sidechain as indicated above. Thr (T) 151 was found to be not important for binding either ligand [26], despite the apparent interaction with the EMP1 ligand noted by Middleton et al. and the fall in exposure score from 6 to 3. Though EMP1 has no obvious sequence or structural homology (but see Section 2.2), it suggested that residues may represent a minimum recognition surface for binding to the receptor [26]. That *RLED* and *RLED* showed virtually no reduction in exposure might simply be assumed to suggest that visual inspection is an inadequate procedure, but these are central parts of two of the 5 binding loops noted by many authors including Middleton et al.

The following is a condensed representation of the output from the same sidechain exposure algorithm as used above [19] but now applied to sidechains in the EMP1 peptide agonist [19]. Again, X means a degree of exposure of 10 or more, and again in practice it almost always means 10. The first row of exposure scores is for the set of reported coordinates for the PDB entry 1EBP for just one chain (C) of the EBP1 ligand homodimer, i.e. the second EBP1 chain (chain D) and the coordinates for the whole of the receptor homodimer have been removed. The second row is the same but for chain D. Note that the symmetry of exposure scores is not perfect, but (noting that X is 10 in this case), they are very similar as Fig. 1 suggests. The third row gives scores for the chain C of the whole EBP1 homodimer (chains C and D) but still without receptor coordinates. The fourth row gives scores for the chain D of the whole EBP1 homodimer (chains C and D) but still without receptor coordinates. Again, note the reasonable similarity as the symmetry in Fig. 1 would suggest. The fifth row is for chain A for interactions in the whole ligand-receptor complex (chains A, B, C, D) as reported in PDB entry 1EBP. The sixth and last row is the same but for chain D. As for the individual chains in the EMP1 agonist homodimer, there is evident symmetry though not quite as perfect in this case. Recall that all these do not, and are not intended to, consider conformational changes on ligand binding, rather they reflect interactions between residues of importance as arise in the final overall ligand-receptor complex. The amino acids in bold and underlined are those which are exposed in both the EMP1 monomer and EMP1 homodimer, and for which exposure decreases by 3 or more units. Details specifying what interacts with what are not revealed by this method used alone but note that glutamine (Q) sidechain of EMP1 hydrogen bonds to the glutamine sidechain on other EMP1 strand. C and D represent the two chains in the EMP-1 homodimer structure, corresponding to the notation used in the EMP-1 PDB entry.

```
   TYSCHFGPLTWVCKPQ EMP1 primary structure
(1) X7127909X8832758 Isolated EMP1 one-chain structure (C)from 1EBP
(2) X7106X09X8720467 Isolated EMP1 one-chain structure (D)from 1EBP
(3) 7510783XX9753878 Isolated EMP1 homodimer structure (C+D)from 1EBP(C)
(4) 7620693XX9652677 Isolated EMP1 homodimer structure (C+D)from 1EBP(D)
(5) 9220X51X96672X3X EMP1 dimer plus receptor structure in 1EBP (C)
(6) X330852XX4661749 EMP1 dimer plus receptor structure in 1EBP (D)
```

The implications in terms of binding energy and importance for agonist activity are discussed later below. For the moment, it may be noted that the trend in decreasing exposure on EMP1-receptor binding is consistent with the observations of Middleton et al. [19-21] who noted Phe 93, Met 150, and Phe 205 of the EBP interact with Tyr (Y) 4, Phe (F) 8, Trp (W) 13, and Cys (C) 15 of the EMP1 peptide to form the main hydrophobic core of the interaction between EBP and EMP1. However, the decrease of exposure of tryptophan W and cystine (C) on binding is significantly less, and proline seems significantly buried on binding, at least by the criteria of the method in Ref [19]. For design purposes discussed below, it is noteworthy that EMP1 interacts with the receptor predominantly through its sidechains. A possible important exception is interaction with EMP1 hydrogen bond with the sidechain of Thr 151 in the receptor it is part of a network of mostly backbone NH and CO hydrogen bonds formed with main chain

atoms of a type I β-turn in EMP1 [26]. Two hydrophobic amino acids, Tyr4 and Trp13, appear essential for mimetic action, and aromatic residues appear to be important at these sites [19]. These findings are consistent with the previously reported X-ray crystal structure of EMP1 complexed with the extracellular domain of the EPO receptor (EPO binding protein; EBP). In efforts to define the structural elements required for EPO mimetic action, a 13 amino acid peptide was identified which possesses mimetic properties and contains a minimal agonist "epitope". Interestingly the ability of their synthetic peptide to act as a mimetic capable of the induction of EPO-responsive cell proliferation appeared to reside within just one residue, equivalent to position Tyr 4 of EMP1, at least when present in a sequence that included a cyclic core peptide structure [19].

3.3 Extending the Data Set: B. Application of Dwyer's Theory

This can be seen as a second step to establishing the basic format, and importantly length, of an alignment frame for training (Section 3.1). Arguably, it does not matter that Dwyer's idea is somewhat controversial and while it may apply in some cases of peptide-protein or protein-protein interactions, it may not apply in others. This is because establishing the format of the alignment frame is a matter of considering the ability to embrace all or most potentially relevant peptides with some similarity rather than relying on their individual validity. Using the standard sequence search and homology tool BLASTP, EMP1 as query TYSCHFGPLTWVCKPQ shows matches with sections of sequence of a variety of proteins of a variety of species, with no obvious significance. Focusing on human proteins, FGPLTW occurs in Immunoglobulin heavy chain junction region, PLTWVC in Plexin B1, SCHFGP in purinergic receptor P2X, ligand-gated ion channel, and with partial matches, PLTWVCK in NADH dehydrogenase subunit 4. With partial matches, leading examples are GPEAWGVCKPQ in the ZYG homologue, and YSCHY-LLTW in immunoglobulin superfamily member 1, and there were weak homologies noted in cytokine and growth factor receptors.

Fig. 2 used the standard sequence alignment tool Clustal Omega to align human EPO (1EPO), its receptor (1EER) on the rationale of Dwyer [31] (see Theory and Methods Section 2). It shows the web page output that can be directly read by knowledge gathering tools [2,27-30] which were helpful, but not essential, here. Over the whole sequences, there is 19% identity between EP and its receptor (same amino acid residues '*') for EPO, and 14% for the receptor. Over the whole sequences, for identical sidechains '*' or highly conserved sidechain properties ':', this rises to 32% for EPO and 23% for the receptor. Note that the output did not include EMP1, which was added subsequently by hand, since alignment is not reliable between such peptides and a reasonably large protein. Even so, the homology match between EP0 and EPO receptor (EPR) and the receptor remains, as one would expect, of a weaker order, but not entirely unpersuasive noting that it lies in a crucial EMP1 and EPO binding region. The first three in the list below are standard characters in Clustal Omega output. Though Dwyer represents a controversial approach, this does seem a possible case in which it does apply although many of the sidechains that would appear to be important in binding correspond to a relatively exposed loop for EMP1 in the ligand-receptor complex. Nonetheless, the evolutionary connection, and information that it may provide on the recognition surface, may be valuable.

'*' - identical residues in EPO and the receptor.

':' – residues very similar in properties by Clustal Omega criteria.

'.' – resides weakly similar in properties by Clustal Omega criteria.

□ (boxes) – Points where arguably the Clustal Omega alignment could be improved or similarities weighted more highly, – either by considering any charged residue replacing a charged residue as a conservative substation, or a single insertion/deletion.

Bold font – residues in EMP1 peptide identical to at least one of EPO and EPO receptor.

Italic underlined font – residues in EPO receptor EER that are at least weakly interacting with the EMP1 peptides.

Turquoise in EMP1 – residues in 5 loops of EPO receptor EER that are closely interacting with EMP1 peptides in both EMP1 chains.

Turquoise in EER – residues in EMP1 peptides that are closely interacting with 5 loops of EPO receptor EER (both chains).

Yellow – range of peptides in EPO sequence as proposed by authors spanning helices C-1 and C-2 of EPO and conserved in human, mouse and rat EP.

Green – range of shortest peptide proposed by authors spanning helices C-1 and C-2 of EPO and conserved in human, mouse and rat EPO.

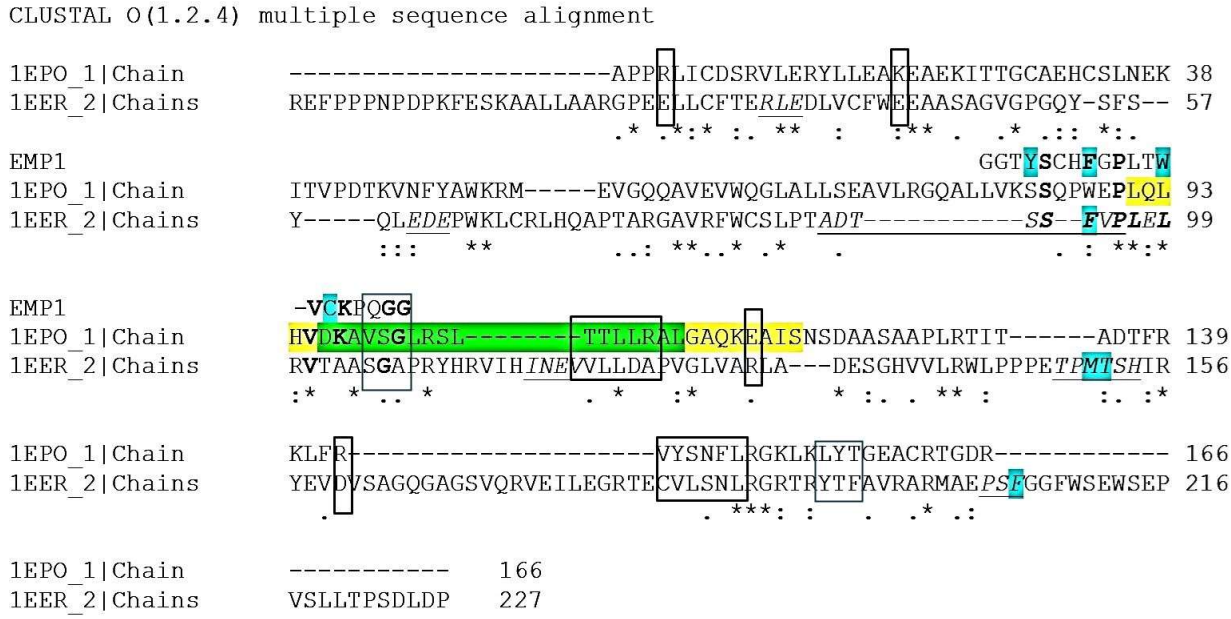

Figure 2. Standard Clustal Omega Printout of Alignment of Erythropoietin, Receptor, and EMP1 Peptide.

3.4 Extending the Data Set: C. Further Studies and Summaries for Specified Peptides

Some studies on arguably related peptides that were experimentally inactive (e.g., Ref. [32]) were subsequently found but did not change these results. Table 2 summarizes previous and further findings, including in terms of exposure and of free energy calculations using the Robson-Platt potential functions [33]. This takes some account of free energies of interactions as well as potential energies. Considering all the approximations for the binding processes implied in the crystal structure, estimates are only to the nearest tenth of a kcal/mole. Estimates of the free energy of interaction between sidechains or whole residues involved in ligand-receptor interaction are ideally of more interest than just contacts or numbers of atoms in the vicinity [34-38] because they better express the strength and importance of the interactions. The peptide considered is the 16-residue peptide since the GG termini remain disorganized in solvent in the crystal and are assumed to be in dynamic random coil state both before and after binding. The experimental free energy of EMP1 binding to the receptor is approximately -7 kcal/mol according to experimental binding data [19] but that is the binding of the EMP1 homodimer to the receptor homodimer, so care must be taken in specifying what is meant by free energy contributions per residue. In Table 2, the Exposure Score results are expressed for the interactions of each sidechain twice, in effect because, due to the homodimerization of the synthesized peptide agonists, the researcher benefits from twice the value that he or she might expect for each mole of residue included in the synthesis of the EMP1 monomer. The physical contribution per residue in terms of kcal/mol is per residue, so half that value. Changes in electrostatic interactions can change greatly from the crystal used in structure determination to the biological conditions and are an adjustable contribution in the potential functions used. This is commonly the case for any choice of potential functions except for the most prolonged simulations and even then, convergence of entropy is extremely difficult. In general, there is a recognized need for rescaling the free energies, especially when estimating free energies of a per residue basis [34,36-38]. Authors note that the scaling problem confounds protein free energy calculations, especially on a per-residue basis for free energy contributions. When possible, it is typical to rescale the free energy values to fit experimental binding data or results

for very accurate simulations on a high-performance computer [34]. Recall that experimental binding data was available in this case.

Description	T	Y	S	C	H	F	G	P	L	T	W	V	C	K	P	Q
Change in Exposure Score of EMP1 (1EDP chain C) on binding	2	-3	1	0	3	-3	-2	0	-1	-3	-1	2	-1	2	-4	2
Change in Exposure Score of EMP1 (1EDP chain D) on binding	3	-1	1	0	2	-4	-1	0	0	-5	0	1	-1	1	-3	2
ΔF kcal/mol interaction EMP1 (1EDP chains C+D) on binding (per residue)	1.9	-1.5	0.8	0.0	1.9	-2.7	-1.2	0.0	-0.4	-1.9	-0.4	0.8	-0.8	1.2	-2.7	1.5
Hydro-phobic core		y				y					y					
Important for activity		y									y					
Conserved in erythropoietin alignment (Section 3.2)			y					y	y			y		y		
Same sidechain class in erythropoietin alignment (Section 3.2)		y	y			y		y	y		y	y		y		
Conserved in erythropoietin receptor alignment (Section 3.2)			y			y		y	y			y				
Same sidechain class in receptor erythropoietin alignment (Section 3.2)			y			y		y	y		y	y				

Table 2. Change in exposure scores of sidechains in going from isolated EMP1 homodimer as ligand to the full ligand-receptor complex.

Negative free energies for residue interactions with the environment tend to lie in the approximate range -1 to -3 kcal/mol. Such is expected to be the case for hydrophobic interactions, which is the case except for valine V in EMP1. Because V seems to be important for recognition one should not of course immediately eliminate it from any design. Theoretically, it may, for example, be important for intramolecular (intrachain) interactions that determine the local conformation of the backbone. Valine prefers a β-conformation and in EMP1 the amine NH and carbonyl CO groups of the backbone form hydrogen bonds with the CO and NH respectively of the histidine H backbone. This may be important for the conformational details of the intramolecular (intrachain) disulfide bridge between the two cystine C residues.

3.5 Extension with Alignments

Extending the data with new peptide entries typically introduces new deletions and hence a new alignment frame on which to train but does not alter the principles and overall procedure. Relevant information for extending and aligning the data for the core 111 peptide entries was obtained from Refs [12-21,32,40-45] and the standard GenPept and PDB databases, combined with Dwyer's Theory. For PDB entry 1EBA_C for EPO mimetics peptide 33, recalling that the X indicates any residue in the original study, GGTXSCHFGPLTWVCKPQGG, bound to the receptor, the X was replaced by a blank in the data described used here to mean any residue can occur there since a blank or the entry 'unknown' means an unknown value in the software used. It also increases the diversity of associations (measured as mutual information) compared with Table 1, as shown in Table 3. Alignment is helped by making

variations in the placement of insertions by optimization of information content (Table 3) and by predictability by DiracSmash. Such association tables are of direct value in defining a consensus structure such as YXCXXGPXTWXCXP, where X returns to representing amino acid residues but in this case means several deduced from source data, not any amino acid. Though these are of course really "unknowns" as far as activity or inactivity is concerned, it is a common tactic in small organic non-peptide drug discovery as discussed in Ref [8].

information	event1	event2	event3	event4
2.61	25:=L	26:=R	27:=S	Active:=yes
2.59	22:=S	25:=L	26:=R	Active:=yes
2.59	20:=A	26:=R	27:=S	Active:=yes
2.59	20:=A	25:=L	27:=S	Active:=yes
2.57	16:=H	21:=V	27:=S	Active:=yes
2.57	16:=H	30:=T	33:=R	Active:=yes
2.57	18:=D	21:=V	34:=A	Active:=yes
2.57	16:=H	32:=L	34:=A	Active:=yes
2.57	16:=H	30:=T	32:=L	Active:=yes
2.57	16:=H	30:=T	34:=A	Active:=yes
2.57	31:=L	34:=A	35:=L	Active:=yes
2.57	18:=D	33:=R	34:=A	Active:=yes
2.57	18:=D	30:=T	31:=L	Active:=yes
2.57	16:=H	27:=S	35:=L	Active:=yes
2.57	30:=T	31:=L	34:=A	Active:=yes
2.57	31:=L	32:=L	34:=A	Active:=yes
2.57	31:=L	33:=R	34:=A	Active:=yes
2.56	27:=S	30:=T	34:=A	Active:=yes
2.56	30:=T	32:=L	33:=R	Active:=yes
2.56	27:=S	31:=L	34:=A	Active:=yes

Table 3. Mutual information in natural logarithmic units for patterns found in the extended small collection of peptides of interest for developing erythropoietic peptides.

Table 4 shows the origins of 32 peptides contributing to the final core 111 that include the effects of deductions based on Sidechain Exposure changes on peptide-protein binding including the method of Dwyer which are represented by identifiers 'a'-'n' in the first column. The row with identifier 'o' represents the form identified by the first step of unsupervised data mining, which was also predictive as active by partially supervised DiracSmash, and the rows 'c', 'd' etc. represent computer "experiments" modifying the sequence and making predictions. The full core data used for the above predictions is still sparse at 111 peptide records, 50 randomly selected being used for training and the remaining 61 for testing predictions, in the manner discussed in Section 3.6 below.

The assumption that the weakly homologous receptor 'h' was to be inactive was strengthened based on simple preliminary binding calculations based on Exposure Scores [30] with free energy interactions estimated with potential functions [36]. However, there seems little doubt that this contains some highly conserved residues associated with the EPO binding site, especially the segment SFVPLE, and more detailed analysis of the interactions with EPO over several species gave support to the interactions discussed above. Recall that the data, 'x' replaced '-' with the same meaning, i.e. that there is one or more deletions in the alignment that again have no impact physically. The residues on either side are directly joined. The sequences with identifiers 'a'-'r' in Table 2 are as follows.

(a) EMP1, residues important for activity in bold.

(b) EMP16 [18].

(c) EMP17 [18].

	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6
a	-	-	G	G	T	**Y**	**S**	C	*H*	**F**	*G*	**P**	**L**	*T*	**W**	-	**V**	C	**K**	P	Q	G	*G*	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	G	G	T	**Y**	**S**	C	*H*	**F**	*G*	**P**	**L**	*T*	W	-	**V**	C	**K**	P	Q	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
c	-	-	-	-	T	**Y**	**S**	C	*H*	**F**	*G*	**P**	**L**	*T*	W	-	**V**	C	**K**	P	Q	G	*G*	-	-	-	-	-	-	-	-	-	-	-	-	-
d	-	-	-	-	T	**Y**	**S**	C	*H*	**F**	*G*	**P**	**L**	*T*	W	-	**V**	C	**K**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
e	-	-	G	G	L	**Y**	A	C	*H*	M	*G*	**P**	M	*T*	W	-	**V**	C	G	P	-	-	-	-	L	R	G	-	-	-	-	-	-	-	-	-
f	A	L	L	V	K	S	**S**	Q	P	W	E	**P**	**L**	Q	L	H	**V**	D	**K**	A	V	S	*G*	-	L	R	S	L	T	T	L	L	R	A	L	-
g	-	-	-	-	-	-	-	-	-	W	E	**P**	**L**	Q	L	H	**V**	D	**K**	A	V	S	*G*	-	L	R	S	L	T	T	L	L	R	A	L	G
h	P	T	A	D	T	-	**S**	S	-	F	V	**P**	**L**	E	L	R	**V**	T	-	A	-	S	*G*	A	P	R	Y	H	R	-	V	I	H	-	I	N
i	-	-	G	G	L	**Y**	A	C	*H*	M	*G*	**P**	I	*T*		-	**V**	C	Q	P	-	-	-	-	L	R	-	-	-	-	-	-	-	-	-	-
j	-	-	G	G	T	**Y**	**S**	C	*H*	F	*G*	**A**	**L**	*T*	W	-	**V**	C	R	P	Q	R	*G*	-	-	-	-	-	-	-	-	-	-	-	-	-
k	-	-	-	V	K	-	S	C	*H*	**F**	*G*	**P**	I	*T*	A	-	**V**	C	**K**	-	Q	S	*G*	-	L	R	G	K	-	-	-	-	-	-	-	-
l	-	-	-	-	-	**Y**	**S**	s	*H*	**F**	*G*	**P**	**L**	*T*	L	-	**V**	s	**K**	A	Q	S	*G*	-	L	R	S	K	-	-	-	-	-	-	-	-
m	-	-	-	-	-	**Y**	**S**	s	*H*	**F**	*G*	**P**	**L**	*T*	L	-	**V**	C	**K**	A	Q	S	*G*	-	L	R	S	K	-	-	-	-	-	-	-	-
n	-	-	-	-	-	**Y**	**S**	C	*H*	**F**	*G*	**P**	**L**	*T*	L	-	**V**	s	**K**	A	Q	S	*G*	-	L	R	S	K	-	-	-	-	-	-	-	-
o			G	G	X	T	X	C	H	X	V	**P**	X	T	W	H	**V**	C	T	P	X	X	X	A	L	R	S	X	T								
p			G	G	X	Y	X	S	H	X	V	**P**	X	T	W	H	**V**	S	T	A	X	X	X	A	L	R	S	X	T								
p			G	G	X	Y	S	S	H	X	V	**P**	X	T	W	H	**V**	S	T	A	X	X	X	A	L	R	S	X	T								
p			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	C	T	P	X	X	X	A	V	R	S	X	T								
p			G	G	X	T	X	C	R	X	V	**P**	X	T	W	H	**V**	C	T	P	X	X	X	A	L	R	S	X	T								
p			G	G	X	T	X	C	K	X	V	**P**	X	T	W	H	**V**	C	T	P	X	X	X	A	L	R	S	X	T								
p			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	C	T	P	X	X	X	A	V	R	S	X	T								
p			G	G	X	T	X	S	H	X	V	**P**	X	T	F	H	**V**	S	T	P	X	X	X	A	V	R	S	X	T								
p			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	S	T	P	X	X	X	A	V	R	S	X	T								
p			G	G	X	T	X	S	H	X	V	**P**	X	T	F	H	**V**	C	T	P	X	X	X	A	V	R	S	X	T								
q			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	C	T	P	X	X	X	A	V	R	S	X	V								
q			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	C	T	P	X	X	X	A	V	R	S	X	K								
q			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	C	T	P	X	X	X	A	V	R	S										
q			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	C	T	P	X	X	X	A	V	R	S										
q			G	G	X	T	X	C	H	X	V	**P**	X	T	F	H	**V**	C	T	P	G	G	G	A	V	R	S										

r			G	G	X	Y	S	S	H	X	V	**P**	X	T	W	H	I	S	T	A	X	X	X	A	L	R	S	X	T											
r			G	G	X	Y	S	S	H	X	V	**P**	X	T	W	H	L	S	T	A	X	X	X	A	L	R	S	X	T											
r			G	G	X	Y	S	S	H	X	V	**P**	X	T	W	H	X	S	T	A	X	X	X	A	L	R	S	X	T											
r			G	G	X	Y	S	S	H	X	V	X	X	T	W	H	**V**	S	T	A	X	X	X	A	L	R	S	X	T											

Table 4. Compromise Peptide Design. Some key peptides are discussed as data (a-n) and guides to an erythropoietic peptide.

(d) EMP24 [18].

(e) EMP35 [18].

(f) EPO natural hormone.

(g) Identification of critical residues in EPO peptides and related studies [19].

(h) EPO receptor in the binding region.

(i) Peginesatide-like peptide.

(j) Pegmolesatide -like peptide.

(k) EPO-receptor-based model considered inactive by conformation imposed by disulfide bond.

(l) Compromise peptide considered active.

(m) Compromise peptide considered inactive

(n) Compromise peptide considered inactive.

(o) Automatic proposal for an active erythropoietic peptide by the “Large Probability Model” (unsupervised data mining followed by supervised data mining by DiracSmash [26]) as described in the text. LR/(1+LR) = 78%. Simplified predictive model gives accuracy=91%, sensitivity=82%, specificity=90%, Diagnostic Odds Ratio DOR = 64.9 at a threshold of 45% established for the “training” set.

(p) “Experimental” modifications to the automatic proposal for which the simplified model gave reasonable predictions LR/(1+LR) = 75-78%. Simplified predictive model gives accuracy=84-89%, sensitivity=82%, specificity=83-88%, Diagnostic Odds Ratio DOR = 34.6-53.8.

(q) “Experimental” modifications to the C-terminus LR/(1+LR) = 77%. Simplified predictive model gives accuracy=77-84%, sensitivity=80-82%, specificity=74-83%, Diagnostic Odds Ratio DOR = 20.3-34.6. The removal of the final two residues treated as XX gave the lowest values in that range, irrespective of where residues 21-23 were replaced by GGG (glycine residues).

(r) Replacing proline P at 12 and valine V 17 had the most deleterious effects giving LR << 1 and DOR << 1, and best predictions for predictive target as Active:=no. Even replacing valine V by isoleucine and leucine for Active:=yes, conservative substitutions, gave an accuracy, sensitivity, and specificity for 36% and 35% respectively.

Predictions reported below depend on the application of Eqn. 1 with (n+v) simply replaced by the counts of actual and assumed activity without distinguishing actual frequencies n from virtual frequencies, as discussed in Section 2.1. following Eqn. 3. While much of the above is automatable, and this is in progress, there are several choices of peptides and their designation as active or inactive that are particularly subjective and harder to treat in a general algorithm. Since the PDB entry refers to GGT(dby)SCHFGPLTWVCKPQGG where (dby) indicates 3,5-dibromotyrosine in position 4 (denoted DBY in the PDB entry) and is inactive, and some similar sequences found by BLAST for unrelated proteins suggest a serine S in that position, an entry with S replacing (dby) was entered as inactive. The alignment between EMP1 cannot imply a conformational homology because the Q in SQP and D in VDK in erythropoietin EPO, aligned with the two cystines in EMP1, are too far apart to form a disulfide bridge without disrupting the conformations, with the SQP in a highly exposed loop and VDK in an α-helix. This does not

completely rule out an evolutionary origin, though it seems much less plausible, but even if it were the case, that would not by itself appear useful for present purposes. However, it also remains possible that the posited homologous sections of sequence between EMP1 and EPO could both bind to the receptor with significant strength, even though in different conformations or conformational changes. On that basis the compromise peptide k was tentatively hypothesized to be active and l and m to be inactive to help ensure that both, and only both, cystine C residue replacement by serine S residues might be considered by the learning procedure. Serine S is a popular substitution for cysteine in protein engineering (due to comparable structure and similarity of sulfur and oxygen), and the serine sidechains hydrogen bond at a comparable distance but can be broken to permit another conformation.

3.6 Predictions of Activity

The studies involved "jackknifing" in which different peptides are removed from the analogue of the training set and used in the test set. As the term is commonly used in bioinformatics, this would in the present study mean that one peptide of known activity or inactivity is removed from the training set and predictions made from them by all the remaining entries before being returned to the training set. Used more generally, this term means that one or more descriptions of a molecule are selected out for predictions and for comparison of each prediction with the known answer; irrespective of the number in each block being selected and tested, the important point is that they are not included in the training set which provides the information for the predictions. Initially, jackknifing of one peptide at a time was done. However, such tests gave more than 80% sensitivity and 75% specificity, which seemed good for sparse data. The normal concern in drug discovery in such a situation would be that the compounds involved are not entirely independent but generated by similar means, or as progressive refinements by the medicinal chemist [8]. In the converse case, if a compound completely related to the others shows up with activity in a report, one expects to see some account of its distinct origin of the choice. Interdependence is of course self-evident in the present case, as Table 4 shows that the sample consists of groups of peptides that can be considered as homologous. Since the chance of interdependent sequences being used for the training and the test is much higher if just one peptide is taken out as the test case at a time, the main studies were also done using separated "training" on the 50 randomly selected non-obviously related peptides described in Section 3.5 and test sets on the remaining 61 sequences. Despite the high probability of interdependence, sparsity also gives rise to an effect that seems contrary to the above. In principle such studies can help identify the features that are important for activity, though in the case of very sparse data there is inevitably more "granularity" to the descriptions of what makes a sequence active, i.e. they are often somewhat isolated sample points in the space of possible active structures. For example, one might simply note that an acidic sidechain is important at a position but not always, without being able to combine it with activity of other sequences to deduce a richer model.

Despite the above cautions and a variety of related studies, results remained consistently high at around 80% sensitivity and 70%-90% specificity. See Discussion and Conclusions Section 4. Recall that in contrast to the accuracy, sensitivity, etc., the LR/(1+LR) is the accuracy at the point when the accuracy equals sensitivity and specificity representing estimates of predictive power of the training set of 50 sequences that produced the LR. This avoids bias toward positive or negative predictions and does not and does not depend on a threshold that underlies a ROC curve. Remarkably, it remained approximately constant at 74-78 for all the current studies, with the sensitivity and specificity etc. also persistently high. Table 5 shows example DiracSmash [26] prediction outputs for the above for a variety of peptides, such as those identifiers 'o' and 'p', in Table 4. See later below for studies 4 and 5.

Agreement with accuracy etc. would confirm extensibility to the test set of 61 sequences and be promising for future data, but an improvement in prediction for the test set is suspicious as essentially a random fluctuation typical of sparse data. However, similar results were obtained by random selection of train and test sets, including extension to a total of 5000 sequences with the additional 4889 comprising assumed inactive from aligned weakly homologous sequences and some or irrelevant or random sequences, initially trained and tested in similar proportion 50:60 for train-test dataset sizes. For the test set the simplified models gave similar results to the full method using many "tags" with further information derived by data mining "inside" DiracSmash. These models tend to do well for certain classes of problem when the data describes attributes that are somewhat compositional in character [8]. Although attributes here include position, e.g., 9:=H, the meaning of the 9 is only apparent to the user, and to the

Study	LR/(1+LR) %	Best threshold %	Accuracy%	Sensitivity%	Specificity%	Quality% (see text)	Diagnostic Log Odds Ratio ± SE
1. Table 4 (o)	78	45	91	82	90	90	4.17±1.17
2.Table 4(p)), '9':='H' → '9':='R'	78	45	92	82	92	91	4.39±1.19
3. Table 4(p)), '29':='T' → '29':='x', '31':='G', '32':='G', '33':='G'	77	35	77	82	74	76	3.01±1.11
4. See text.	78	40	80	82	78	79	3.20±1.12
5. See text.	75	40	86	82	85	86	3.68±1.14

Table 5. Results of several prediction studies discussed in the text.

algorithm it is not materially different to a compositional description of a compound such as 'number of hydroxyl groups':=3. With a focus on seeking to improve on EMP-1 including reducing side effects, then for amino acid residues at locations in the sequence where predictions were weak or many variations of amino acid residue would do, the residue found in EMP-1 is proposed.

GGTYSC[H/K]FGPT[W/F]HVCK[P/A]VGGLRSG
GGXTXC H---XVPXTW---HVCTP------XXALRSXT (automatic model (o))

The second sequence above is the automatically generated model 'o' in Table 4 where again X represents several possible substitutions. The new proposal represented by the upper sequence above seems less promising than the less specific proposal (by containing several 'X') made by the automatically generated model 'o', although the LR/(1+LR) = 78% from the full DiracSmash calculation provides some justification for its further consideration, with results as shown for Study 4 in Table 5. There was significant improvement in quality of prediction by the simplified model when substituting the two cystines C by serine S, which likely loses the benefit of the stability of the covalent intramolecular disulfide bridge, but which might simplify preparation of pure peptide. However, the full model gave a reduced LR/(1+LR) = 73%, in that sense disagreeing with the benefits of the S for C substitutions for increased potency. The following peptide gave the prediction results shown as Study 5 in Table 5.

GGTYSS[H/K]FGPT[W/F]HVSK[P/A]VGGLRSG

3.7 Comparative and Advanced Studies

As a final step in this early stage of investigation, meaning just before the investigation has facilitated gathering much larger amounts of data (including by synthesis and testing), incorporation of further methods considered as AI, such as Machine Learning and Deep Learning, is valuable. Comparisons can give support to the findings obtained so far and, since methodologies may only partially overlap, there is the possibility of capturing further information of value. Essentially the same methods are used here as in Refs [8,11,47,49]. See Table 6 that refers to studies 1-5 at the end of Section 3.6. DL is typically considered as not well-suited to small datasets, although there have been efforts to explore and overcome the problem (e.g., Ref [50]), and there are important exceptions and techniques that make it workable in specific scenarios (especially in vision, language, and audio tasks, but not drug development). Recognized problems [50] are that DL can have millions of parameters and so tends to overfit small data.

Study	GB with TEI, sensitivity	GB with TEI, specificity	GB without TEI, sensitivity	GB without TEI, specificity	ML, sensitivity	ML, specificity	DL, sensitivity	DL, specificity
1	82	90	70	74	71	74	61	68
2	82	92	75	80	75	80	68	71
3	74	76	72	74	75	75	-	-
4	82	78	55	58	-	-	-	-
5	82	85	54	57	-	-	-	-

Table 6. Comparison of Glass Box GB Approaches with Machine Learning and Deep Learning.

4 Discussion and Conclusions

4.1 Development of Peptidomimetics

The method of selecting peptides for consideration is indicated in Section 2.5 and discussed in Results Section 3. It clarifies the importance and nature of a glass box approach. There it was noted that a major query of interest in the study was active:=yes, 3:=G, 4:=G, 5:=T, 6:=Y 8:=C, 9:=H, 9:=K, 12:=P, 14:=T, 15:=W, 15:=F, 15:=W, 16:=H, 17:=V, 18:=C, 20:=P, 20:=A, 21:=V, 21:=G, 22:=G, 23:=G, 25:=L, 26:=R, 27:=S, 29:=x, 34:=A. Here it was with '12':='P', and '17':='V' as interdependent attributes and the rest as independent attributes. But even without that consideration, and while some care is required to interpret to the positions of residues in the light of the deletions, this is a clear guide to what to consider next, including ultimately synthesis, as it was with small non-peptide compounds [8]. Such a query is generated by unsupervised data mining, and the role of the query (in DiracSmash) is to do further mining focused by the query and confirm the description by, primarily, a high Likelihood Ratio (or reject it by a low one). The risks of being misguided are inevitably higher when data is very sparse, but the general idea has been supported by the study on small organic compounds where more data was available both for the step analogous to training and for blind testing [8]. The primary difference in the present study is that the nature of peptides allows the study to be structural, not just compositional, and the ability to use weaker evidence as "hunches" in a manner facilitated by the relevance of bioinformatics tools.

Sensitivity to hydrolytic peptidases and acid hydrolysis in the gut are a major issue and making non-biological modifications is the most plausible next step, and with more extreme modifications may represent an intermediate step towards a small organic compound, but this requires different techniques and will be discussed elsewhere. However, extensive use of D-amino acids has a more direct mapping to the L-amino acid sequence candidate. A *retroinverso* peptide [46], for example, confers resistance to enzymic hydrolysis. The sequence is written backwards (residues in reverse order) but using D (dextro, mirror image) amino acid residues. In the case of EMP1, that gives dextro-(QPKCVWTLPGFHCSY). In a *retroinverso* structure it is conceptually as if the sidechains were in the same position in space, but the peptide groups linking residues, (NH-CO) are reversed (CO-NH). This highlights another good reason for considering sidechain and other interactions. If interactions between receptor-sidechain to peptide-sidechain and receptor-backbone to receptor-sidechain are important, there is a chance of activity, but if the peptide-backbone is involved in strong interactions with the receptor, they will almost inevitably be hydrogen bonding interactions NH…OC, and the location of NH and CO groups in the backbone peptide link have been reversed, making the interaction likely repulsive. Consequently, the following are plausible candidates for synthesis and testing.

dextro-(GSRLGGV[A/P]KCVH[F/W]TPGF[K/H]CSYTGG)
dextro-(GSRLGGV[A/P]KSVH[F/W]TPGF[K/H]SSYTGG)

4.2 Advantages and Limitations of the Approach

There has been no validation by new synthesis and testing in the present study, which has essentially comprised a proposal for study of peptide activity when data is sparse, but as in many compound activity studies (e.g., Ref [8]),

the equivalent of a training and test set has been constructed (Section 3.6). This project suggests that predictions of peptides as active or inactive can be carried out with reasonable predictive capability and good sensitivity given sparse data, which is often the case early in pharmaceutical research. Reasonable predictions obtained were for peptides where the property of active or inactive was known. The fact that a Glass Box partially supervised prediction approach possible to test differing descriptions of peptides of interest (the query), rather than simply predict that a particular compound is active or inactive, turns out to be well suited to peptide design, allowing optimization of designs as well as discovering significantly different structures with similar activity (but differing in other merits). As in previous non-peptide work [8] it can provide a guide as to what to query or synthesize next.

The hunch strategy is intended to increase the odds of success in selecting peptides for a more reliable second phase prediction or experimental selection strategy. Also, this second phase, or a phase following that, will almost certainly consider peptidomimetic modifications for animal studies or because the benefits of such modifications (resistance to gastric acidity and peptidases etc.) are needed for clinical trials. The present study has attempted to make explicit, and ultimately automatable, methods in which a Glass Box AI-style approach has advantages by dealing with sparsity and a formal relationship to inclusion of plausible prior beliefs.

It has been emphasized that Dwyer's theory is controversial, but the approach is such that the overall hunch approach is robust even where it does not hold. First, it is assumed that any peptide deduced by that means is inactive, and as discussed by analogy with pharmaceutical discovery at comparable stages that is much more likely be the case. Second, the important thing is that it has certain similarities in sequence pattern to the active peptides, at least at the level shown in Fig. 2, and it need not be used if there is no relationship at all. The more precise probabilistic description of what constitutes a significant relationship in this kind of situation is arguably a matter needing future work, but in view of the first point, there seems little to be lost by always assuming it to apply provided it is implemented as the maximum alignment, as in Fig. 2. Not least, if the degree of match is statistically very weak, it is more likely to be inactive.

Comment should be made on the quality of the results as judged by sensitivity, specificity etc. because intuitively it seems rather high for a study involving sparse data. Section 3.6 discussed the preferred choice of sampling and separation of "training" and test cases, based on the concern that that the peptides are far from independent. Selecting a peptide for prediction and leaving many close relatives in the datamined set would clearly give predictions that appear, and arguably genuinely are, of good quality. The situation may be close to the norm in drug discovery because compounds are not generated independently, or because they are progressively modified by the medicinal chemist, or some combination of the two [8]. A conclusion might be reached that the quality of the results obtained is thus illusory, and not helpful for therapeutic discovery. As was also argued elsewhere [8], this is not the case. Blocks of data for compounds that are related by some kind of common origin and represent exploration to refine activity are the normal data ecosystem for drug discovery, so good use should be made of it. The kind of study described here is in a sense and in part a continuation of that process.

4.3 Future Work

A preliminary observation worthy of further investigation is that compounds strongly suspected of inactivity when included in the tests as inactive did not greatly change the predictive performance [32]. Continuing to add data of that kind will help clarify the relative merits of different methods of including less direct data. Other approaches to explore use of structural data and predictions from it can be expected to increase the usefulness of the method (e.g., Deep learning for sparse data [50]) and AlphaFold technology [51], along with large language models for peptide sequences [52], extensions of DiracSmash for less perfect real world data [53], and more extensive use of the proposed Q-UEL language with which DiracSmash is compatible [54,55]. As to future peptides of interest, several new classes of natural biologically active peptides are emerging, including many derived by direct translation from small open reading frames (sORFs) originally thought too small to be genes [55]. But not least, erythropoietin has many different actions at diverse receptors distinct from erythropoiesis, so that further peptides derived from studies of the hormone for applications such as neuroprotection, and they face similar challenges of relatively sparse data [56].

The present method ultimately rests on estimates of real probabilities and is not prone to "hallucinate" in the same sense as current AI methods based on learning weights. Since hunches based on weak evidence are an important feature at the start of the process described here, there are obviously opportunities for errors. The researcher is merely choosing a plausible path based on the evidence available. Unlike "black box" AI approaches, however, the basis of the reasoning and the meaning of the probabilities and odds ultimately generated are visible (as the probabilistic Knowledge Element Tags, Section 2.5), auditable, and open to investigation, correction, updating and reuse. Arguably, the AI approach most relevant to the present discussion is AlphaFold [51]. That technology is powerful and there are several ways in which it might be brought into play as steps following the above. However, it is not a *de novo* protein folding approach [57] based on physics. It requires large amounts of data and depends on the conservative nature of protein evolution. The benefits and limitations of AlphaFold have been reviewed by the present author [58] and the points raised appear valid today. A very limited understanding of what will make an active peptide typically means that there is less information to make good use of AlphaFold. It would clearly be very valuable to refine the proposed therapeutic agents by modeling their interactions with the target, but there is obviously a huge advantage in first proposing peptide ligand structures to explore in that way. Researchers use AlphaFold as part of a pipeline, (i) to predict the structure of a disease-related protein, (ii) model how a candidate peptide might interact with that protein and (iii) refine peptide sequences based on predicted interactions. It is especially helpful when an experimental structure for the protein target is not available, and it was available in the present study. It does not at present generate novel peptide drugs "from scratch", which is the primary purpose of the approach described here.

Finally, an obvious future task will be translation of the results of the above methodology into a pharmacophore model for small molecule screening that has beneficial impact for medicinal chemists. It should be kept in mind that this is not the only interest: peptides are occasionally approved and used directly as therapeutics. Novel active peptides can also facilitate the development activity tests for non-peptide compounds without direct consideration of a pharmacophore. However, in the present paper, an obvious question might be how the "Knowledge Element Tags" (Section 2.5) for peptides can be used to aid design of small organic "in a pill" drugs. These represent the formal, persisting embodiments of the knowledge captured by the study, retained alongside other similarly captured knowledge in large Knowledge Representation Stores for medicine and pharmaceutical research [59]. In the present paper they are used in predictions but are already in the canonical form called the Q-UEL language [53-55] that is available for reuse in future work, potentially including processes of automated reasoning. Note that each such tag considered individually is not an estimate in the same sense that inference and prediction overall involve independence assumptions. Each tag represents the use of data when it is sufficient to make exact calculations of useful probabilistic measures as ratios of counts (numbers of observations). The notion of "estimate" is confined to the limit number of counts involved. This does, however, mean that the knowledge contained in these tags is somewhat confined to use in bioinformatics and biotechnological studies. It is, at best, mappable with further studies to a small and likely disconnected representation of a van der Waals, electrostatic, and hydrophobic surface representation that a smaller molecule might similarly possess. However, nothing prohibits capturing more integrated representations that can be captured during a project, such as those in Table 4, which can readily be re-expressed as attributes (descriptors) in the Q-UEL tag form. Similar Q-UEL representation of knowledge in many bioinformatics and pharmaceutical studies, e.g., Refs [2,27-30,55]. Most of that is, nonetheless still of a bioinformatics nature.

Consequently, the most exciting area under investigation in the present larger project is the accumulation of the kinds of Q-UEL Knowledge Element Tags generated from activity data for small organic molecules as in Ref [8], with the specific intent of using these alongside the above tags for peptide data to seek to infer, ultimately automatically, what kinds of description of a small organic molecule might map to those of an active peptide, i.e. what organic molecules might the peptide data imply? Although the term "organic chemistry" is now essentially historical and refers only to the presence of carbon atoms, it is historical for good reason. Through metabolism, biochemistry makes use of a recurrent set of chemical motifs (including, interestingly, those that often turn up in news reports of potential evidence for life on other worlds). The combinatorial consequences of chemically joining these have been explored (e.g., Ref [60]). The motifs can be put together in an enormous combinatorial variety of Markush representations, each with a variety of options that pose challenges for the notion of novel compositions of matter in patent law [60].

Because proteins as drug targets have binding site pockets lined with side chains such as lysine, aspartate, serine and phenylalanine, and in awareness that just 20 kinds of amino acid residues interact to produce folded structures and protein-protein assemblies with a seeming infinite variety of functional structures [60], medicinal chemists intentionally use corresponding motifs even when they do not start from a peptide ligand. These motifs include amines, carboxylic acids, and hydroxyl groups, as well as larger aromatic and heterocyclic groups. Many of the attributes (descriptors) that are available drug discovery data (e.g., Refs [61,62]) and captured on the tags are quite large aromatic and heterocyclic groups to which features of peptides might be fitted, as well as physicochemical data for the molecules in entirety.

It seems unlikely that developments into protein-ligand interactions by advances in approaches like those of AlphaFold [51] will make approaches like the above study redundant, and it certainly it will be essential to bring in knowledge that can be captured and utilized in tag form but beyond the remit of AlphaFold which was primarily designed for protein structure prediction. Admittedly, AlphaFold 3 has moved from that toward direct, high-accuracy prediction of drug-like interactions with proteins [63]. Nonetheless, the huge variety of combinatorial structures for organic molecules in which small changes in chemistry can cause impactful changes in atomic position [60] must mean that data for specific areas is rather too sparse and inappropriate for reliable training. AlphaFold 3 model is described as having a substantially updated diffusion-based architecture capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues used in peptidomimetics. However, it remains that AlphaFold-like methods learn patterns from biological data and do not generalize well to arbitrary synthetic chemistry space. AlphaFold 3 answers questions such as “Given this protein and this ligand, what complex structure is plausible?”. Discovery of small organic drugs requires generating novel molecules by exploring vast chemical space and optimizing multiple properties, and drug design requires conceiving molecules that do not yet exist and, in order to be patented, must be as different as possible from molecules that are natural or patented. AlphaFold-style models are not directly related to obtaining binding affinity or handling dynamic motion and kinetics. Real drug design must optimize solubility, suitability, interactions with metabolism and membrane permeability. It has no notion of ADMET, the key pharmacological properties of absorption, distribution, metabolism, excretion and toxicity issues that a drug must satisfy to win approval from bodies such as the FDA. Nonetheless, it seems clear that that the more specialized tools like AlphaFold can provide important elements of knowledge that contribute to the drug discovery process and potentially do so by the kinds of Knowledge Element Tags discussed here. The argument that the approach can do this is discussed with examples in Refs [2,8,26-30,53,54,54,59,64].

Ethics, Conflicts-of-interest and Data Availability Statement

This study is primarily an example of a Glass Box Machine Learning approach applied to data already publicly available. No animals or humans were involved. The software used is proprietary to Ingine Inc., but the mathematical basis and algorithms used have been published extensively as described in Section 2.1. See especially Refs [26,28]. A variety of sequences were explored as the data was progressively extended, as described in the text, but the main core of 111 are derivable from public data sources. These 111 are available on request from Ingine Inc. The results reported are novel and the author’s own work. No human or animal subjects were involved. The author BR is a cofounder of Ingine Inc. and holds stock and is a Senior Editor of Journal of Biomedical Informatics and AI of Concetta Press. He is also CEO of The Dirac Foundation, Oxfordshire UK, a non-financial organization with an open policy on publishing to promote the ideas of Nobel Laureat Paul A. M. Dirac for human and animal medicine.

References

1. B. Robson, The design of biologically active polypeptides. CRC Crit Rev Biochem. 1983;14(4):273-96. doi: 10.3109/10409238309102796.

2. B. Robson, Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus, Computers in Biology and Medicine, published online 26 February 2020, 103670, (2020).

3. N. Nissan, M. C. Allen, D. Sabatino, and K. K. Biggar, Future Perspective: Harnessing the Power of Artificial Intelligence in the Generation of New Peptide Drugs. Biomolecules. 14, 10, 1303, (2024).

4. M. Goles, A. Daza, G. Cabas-Mor , L. Sarmiento-Varón, J. Sepúlveda-Yañez, H. Anvari-Kazemabad, M. D. Davari, R. Uribe-Paredes, A. Olivera-Nappa, M. A. Navarrete, and D. Medina-Ortiz, Peptide-based drug discovery through artificial intelligence: towards an autonomous design of therapeutic peptides, Briefings in Bioinformatics, 25, 4, (2024).

5. S. Zhai, T. Liu, S. Lin, D. Li, H. Liu, X. Yao, and T. Hou, Artificial intelligence in peptide-based drug design, Drug Discovery Today, 30, 2, 104300, (2025).

6. L. Wang, X. Fu, X. Ye, T. Sakurai, X. Zeng and Y. Liu, PKAN: Leveraging Kolmogorov-Arnold Networks and Multi-modal Learning for Peptide Prediction with Advanced Language Models, IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2025.3561846, (2025).

7. D. Carou, A. Sartal, and J. P. Davim, and (Eds), Machine Learning and Artificial Intelligence with Industrial Applications: From Big Data to Small Data, Springer, (2022).

8. B. Robson and R. Cooper, Glass Box and Black Box Machine Learning Approaches to Exploit Compositional Descriptors of Molecules in Drug Discovery and Aid the Medicinal Chemist, ChemMedChem, (Wiley), e202400169, (2024).

9. Frenkel, D., Clark, D.E., Li, J. et al. PRO_LIGAND: An approach to de novo molecular design. 4. Application to the design of peptides. J Computer-Aided Mol Des 9, 213–225 (1995).

10. M. Pratt, Causes, Symptoms and Management of Anemia, American Medical Publishers, (2022). ISBN :9781639271788.

11. B. Robson and O.K. Baek, Glass Box machine learning for retrospective cohort studies using many patient records. The complex example of bleeding peptide ulcer, Computers in Biology and Medicine, 108085, (2024)

12. S. Qureshi, R.M. Kim, Z. Konteatis, et al, Mimicry of erythropoietin by a nonpeptide molecule, Proceedings National Academy of Sciences U.S.A., 96, 2112156-12161, (1999).

13. K. Babaoglu, A. Simeonov, J.J. Irwin et al., Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens against β-Lactamase, Journal of Medicinal Chemistry,51, 8, 2502-2511, (2008).

14. R. P. Bissonnette, D. Rungta, A. R. Hudson, et al. NON-Peptidyl Small Molecule EPO Receptor Agonists Are Potent Pathway Selective Inducers of Erythropoiesis. Blood,116 (21): 1565 (2010).

15. F. Guarnieri, Designing a small molecule erythropoietin mimetic. Methods in Molecular Biology, 1289, 185-210, (2015).

16. G. Li Petri, S. Di Martino, Simona, and M. De Rosa, Peptidomimetics: An Overview of Recent Medicinal Chemistry Efforts toward the Discovery of Novel Small Molecule Inhibitors, Journal of Medicinal Chemistry, 65, 11, 7483-7425, (2022).

17. N. C. Wrighton, F. X. Farrell, R. Chang, et al. Small peptides as potent mimetics of the protein hormone erythropoietin, Science, 26;273, 5274, 458-64, (1996).

18. N. C. Wrighton, P. Balasubramanian, F. P Barbone, et al., Increased potency of an erythropoietin peptide mimetic through covalent dimerization. Nature Biotechnology. 15, 12, 1261-1265, (1997).

19. D. L. Johnson, F. Farrell, F. P. Barbone, F.J. McMahon, J. Tullai,K. O. Hoey, N. Livnah, C. Wrighton, S. A. Middleton, D. A. Loughney, E. A. Stura, W. J. Dower, L. S. Mulcahy, I. A. Wilson,| and L.K. Jolliffe, Identification of a 13 Amino Acid Peptide Mimetic of Erythropoietin and Description of Amino Acids Critical for the Mimetic Activity of EMP1, Biochemistry, 37, 3699-3710, (1997).

20. O. Livnah, E.A. Stura, D.L. Johnson, et al., Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 A., Science, 273, 464-471, (1996).

21. S. A. Middleton, F. P. Barbone, D. L. Johnson et al, Shared and Unique Determinants of the Erythropoietin (EPO) Receptor Are Important for Binding EPO and EPO Mimetic Peptide, Journal of Biological Chemistry, 274, 20, (1999).

22. J. Randal J. FDA panel scrutinizes safety of anti-anemia drugs. Journal of the National Cancer Institute, 96, 1061, (2004).

23. Editorial (unnamed), Erythropoietin analogues: an unnecessary class of drugs, Lancet Oncology, 9,2,81,(2008).

24. B. Robson, Doppelganger Proteins as Drug Leads, B. Robson (1996), Nature Biotechnology, 14, 892-893, (1996).

25. S. French and B. Robson, What is a Conservative Substitution?, J. Mol. Evolution, 19, 171-175, (1983).

26. B. Robson, B. and S. Boray, Studies in the Extensively Automatic Construction of Large Odds-Based Inference Networks from Structured Data. Examples from Medical, Bioinformatics, and Health Insurance Claims Data, Computers in Biology and Medicine, 95:147-166 (2018).

27. B. Robson, COVID-19 Coronavirus Spike Protein Analysis for Synthetic Vaccines, a Peptidomimetic Antagonist, and Therapeutic Drugs, and Analysis of a Proposed Achilles' Heel Conserved Region to Minimize Probability of Escape Mutations and Drug Resistance, Computers in Biology and Medicine, 121, June 2020, 103749 (2020)".

28. B. Robson, Bioinformatics studies on a function of the SARS-CoV-2 spike glycoprotein as the binding of host sialic acid glycans, Computers in Biology and Medicine, 122, July 2020, 103849, (2020).

29. B. Robson, B., The use of knowledge management tools in viroinformatics. Example study of a highly conserved sequence motif in Nsp3 of SARS-CoV-2 as a therapeutic target, Computers in Biology and Medicine, 125, Epub August, 103963, (2020).

30. Robson, B., Techniques Assisting Peptide Vaccine and Peptidomimetic Design. Sidechain Exposure in the SARS-CoV-2 Spike Glycoprotein, Computers in Biology and Medicine, 128, Epub November, 104124, (2021)

31. D. S. Dwyer, Protein Receptors Evolved from Homologous Cohesion Modules that Self-Associated and are Encoded by Interactive Networked Genes, Life (Basel), 11,12,1335, (2021).

32. C. Kessler, A. Greindl, B. Breuer, et al., Erythropoietin mimetic compound AGEM400(HES) binds to the same receptor as erythropoietin but displays a different spectrum of activities, Cytokine, 57(2), 226-37, (2011).

33. B. Robson, B. and E. Platt, E, Refined models for computer calculations in protein engineering. Calculation and testing of atomic potential functions compatible with more efficient calculations, Journal of Molecular Biology, 188, 259-281, (1986).

34. B. Robson, J. Li, R. Dettinger, A. Peters, and S. K. Boyer, 'Drug discovery using very large numbers of patents: general strategy with extensive use of match and edit operations', J. Computer Aided Molecular Design, 255, 427–441, (2011).

35. P. S, Klepeis J. L. and D. E. Shaw, Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. Current Opinion in Structural Biology 24, 98-105, (2014).

36. L.J. Williams, B. J. Schendt, Z. R. Fritz ZR, et al, A protein interaction free energy model based on amino acid residue contributions: Assessment of point mutation stability of T4 lysozyme. Technology (Singapore World Science), (1-2), 12-39, (2019).

37. H. Meirovitch, S. Cheluvaraja, and R. P. White, Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding. Curren Protein and Peptide Science, 10, 229–243 (2009).

38. M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, A Role of molecular dynamics and related methods in drug discovery. J Med. Chem. 59, 4035–4061 (2016)

39. D. L. Johnson, F. X. Farrell, F. P. Barbone et al., Amino-terminal dimerization of an erythropoietin mimetic peptide results in increased erythropoietic activity, Chemistry & Biology, 4, 12, (1997).

40. S. A. Middleton, D. L. Johnson, R. Jin et al. Identification of a Critical Ligand Binding Determinant of the Human Erythropoietin Receptor. Evidence for Common Ligand Binding Motifs in the Cytokine Receptor Familj, Journal Of Biological Chemistry 271, 24, 14045-14054, (1996).

41. J. Frank J. McMahon, J. Tullai et al., Mutagenesis studies of the human erythropoietin receptor: Establishment of structure-function relationships, Journal Of Biological Chemistry, 272, 8,4985–4992, (1997).

42. V. Rakesh, J. M. Green, P. J. Schatz, and D. M. Wojchowski, A dimeric peptide with erythropoiesis-stimulating activity uniquely affects erythropoietin receptor ligation and cell surface expression, 44,8,765-, Experimental Hematology, (2016).

43. J. Goldberg, Q. Jin, Y. Ambroise, et al., Erythropoietin Mimetics Derived from Solution Phase Combinatorial Libraries, Journal of the American Chemical Society, 124(4), (2001).

44. O. Livnah,O., D. L. Johnson, E. A. Sutra et al, An antagonist peptide-EPO receptor complex suggests that receptor dimerization is not sufficient for activation, Nature Structural Biology 5, 993-1004, (1998).

45. C. P. Holmes, Q. Yin, G. Lalonde, P. J. Schatz, D. Tumelty, B. Palani, and G. Zemede, Peptides that bind to the erythropoietin receptor, US Patent US7528104B2,(2009).

46. J. Rai, Peptide and protein mimetics by retro and retroinverso analogs. Chem Biol Drug Des.;93, 724-736, (2019).

47. B. Robson, Clinical and Pharmacogenomic Data Mining: 3. Zeta Theory As a General Tactic for Clinical Bioinformatics" Journal of Proteome Research. (Am. Che. Soc.), 4, 2, 445-455,(2005).

48. B. Robson and O.K. Baek, Use of a theory of expected information for sparse data and adverse events in clinical trials and other biomedical studies, Information Sciences, 680, 121027, (2024).

49. S. A. Ali, M.I. Hassan, A. Islam, and F. Ahmad A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and unfolded states. Current Protein and Peptide Science, 15, 5, 456-76, (2014).

50. L. Brigato and L. Iocchi, A Close Look at Deep Learning with Small Data, arxiv.org/pdf/2003.12843, (2022).

51. E. Ferrario, R. Miggiano, M. Rizzi, and D. M. Ferraris, The integration of AlphaFold-predicted and crystal structures of human trans-3-hydroxy-l-proline dehydratase reveals a regulatory catalytic mechanism, Comput. Struct. Biotechnol. J., 18, 20, :3874-388, (2022).

52. A. Madani, B. Krause, E. R. Greene, et al., Large language models generate functional protein sequences across diverse families. Nature Biotechnology. 41, 8 1099-1106, (2023).

53. B. Robson, S. Boray and J. Weisman, Mining Real-World High Dimensional Structured Data in Medicine and its Use in Decision Support. Some Different Perspectives on Unknowns, Interdependency, and Distinguishability", Computers in Biology and Medicine, 141,105118, (2022).

54. B. Robson, T. P. Caruso, UGJ Balis, Suggestions for a web based universal exchange and inference language for medicine, Computers in Biology and Medicine 43 (12), 2297-2310, (2013).

55. B. Robson, Quantum Universal Exchange Language and Hyperbolic Dirac Nets for Precision Medicine and Drug Design. Proposals with Examples from Mitochondrial Studies. Computers in Biology and Medicine, 117, 103621.

56. B. Cho, S-J Yoo, S. Y. Kim, C-H Lee, et al., Second-generation non-hematopoietic erythropoietin-derived peptide for neuroprotection, Redox Biology, Volume 49, 102223, (2022).

57. B. Robson, De novo protein folding on computers. Benefits and challenges, Computers in Biology and Medicine,143,109242, (2022).

58. B. Robson, Testing machine learning techniques for general application by using protein secondary structure prediction. A brief survey with studies of pitfalls and benefits using a simple progressive learning approach, Computers in Biology and Medicine, 138,104883, (2021).

59. B. Robson, B. and S. Boray, Data-Mining to Build a Knowledge Representation Store for Clinical Decision Support. Studies on Curation and Validation based on Machine Performance in Multiple Choice Medical Licensing Examinations", Computers in Biology and Medicine, 73:71-93, (2016).

60. B. Robson, The Concept of Novel Compositions of Matter. A Theoretical Analysis." Intellectual Property Rights , Intel Prop Rights, 1, 108, (2013), doi: 10.4172/ipr.1000108

61. https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors.html (last accessed May 6 2026)

62. https://github.com/rdkit/rdkit/blob/master/Data/FragmentDescriptors.csv (last accessed May 6 2026)

63. J. Abramson, J., Adler, Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500, (2024).

64. B. Robson, Bidirectional General Graphs for inference. Principles and implications for medicine, Computers in Biology and Medicine, 10,382-399, (2019).