1 Supplementary materials

1.1 Complexity of the UCDR problem

We show that an instance of the decision version of the UCDR problem is NP-complete.

Remark 1. *Given a set of positive (rational) numbers. The problem of determining if there exists two disjoint nonempty subsets whose elements sum up to the same value is NP-complete [Woeginger, G. J., & Yu, Z. (1992). On the equal-subset-sum problem. Information Processing Letters, 42(6), 299-302].*

The problem in Remark 1 was called *"equal subset sum problem"*. Notice that the pair of two subsets in the solution is not necessary a partition (i.e. there may be some elements that are in the original set but are not in either of these two sub-sets).

Theorem 2. *Given a set of points in a n-dimension space where each point was assigned a color either blue or red. The problem of determining if there exists a non-empty dimension subset and a center point such that all blue points are not farther to that center point in comparison to red points (by the L_1 norm in the reduced dimension space) is NP-complete. We call the problem "UCDR decision problem".*

Proof. We will reduce the equal subset sum problem (Remark 1) to a special instance of the UCDR decision problem.
Assume we are given a set of positive rational numbers $A = \{a_1, a_2, \ldots, a_n\}$. We create two blue points $B_1 = (a_1, a_2, \ldots, a_n)$, $B_2 = (-a_1, -a_2, \ldots, -a_n)$ and one red point $R = (0, 0, \ldots, 0)$. We consider the UCDR decision problem of three points B_1, B_2 and R. Suppose that this UCDR decision problem has a solution that includes a dimension subset $I = \{i_1, i_2, \ldots, i_d\} \subseteq \{1, 2, \ldots, n\}$ and a center C.
Now we only consider the reduced space with d dimensions from I. We denote B_1', B_2', and R' as the corresponding points of B_1, B_2, and R respectively in the reduced space.
Let H be the smallest (by volume) L_1 norm ball that has the center C and contains both B_1' and B_2'. Thus B_1' or B_2' (or both) must be on a facet of H, we can assume B_1' is on a facet of H without losing generality. Since H is convex and $R' = (B_1' + B_2')/2$, H also contains R'. But if B_2' is not on the same facet of B_1', then R' will be inside H and thus $d(C, R') < d(C, B_1')$. Therefore, both B_1', B_2' and R' must be on the same facet of H. Let F be that facet, since H is a L_1 norm ball then any point $(x_{i_1}, x_{i_2}, \ldots, x_{i_d}) \in F$ must satisfy an equation that has the form

$$\pm x_{i_1} \pm x_{i_2} \pm \ldots \pm x_{i_d} = s$$

Since $R' = (0, 0, \ldots, 0) \in F$, so s must be 0. Thus we can re-write the equation as

$$\sum_{i_j \in I_1} x_{i_j} - \sum_{i_k \in I_2} x_{i_k} = 0$$

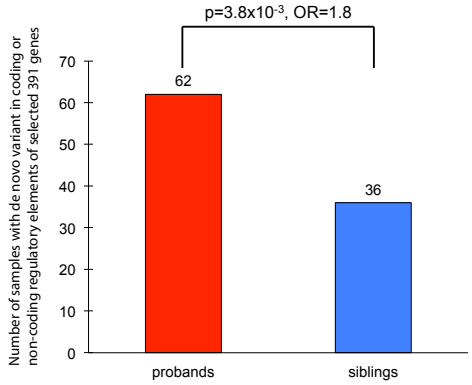
where $I_1 \cap I_2 = \emptyset$ and $I_1 \cup I_2 = I$. Since $B_1' = (a_{i_1}, a_{i_2}, \ldots, a_{i_d}) \in F$ then

$$\sum_{i_j \in I_1} a_{i_j} - \sum_{i_k \in I_2} a_{i_k} = 0$$

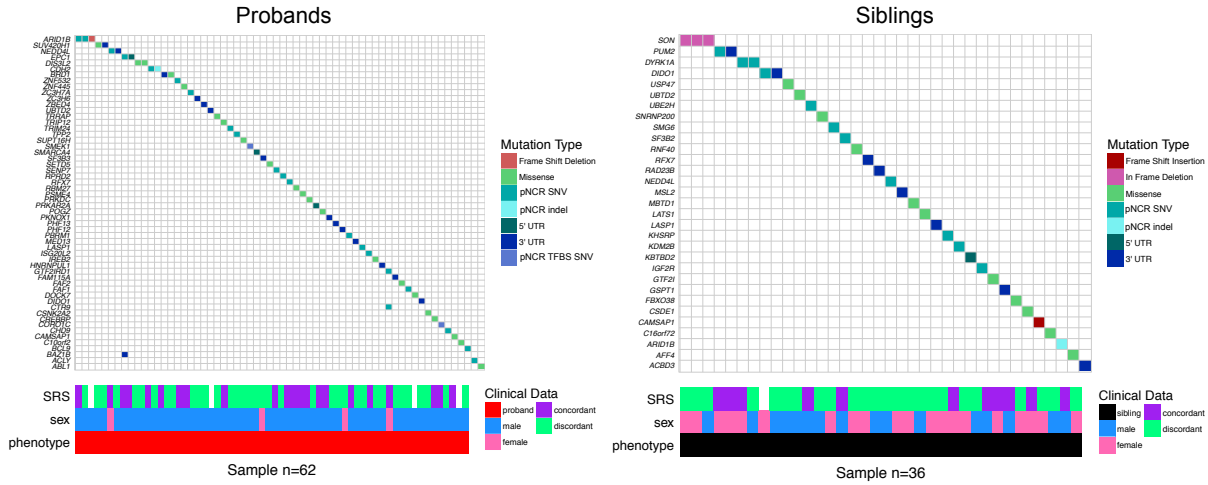
but both a_{i_j} and a_{i_k} are in A that contains positive numbers only so $I_1 \neq \emptyset$ and $I_2 \neq \emptyset$. Therefore, the pair of two sets $A_1 = \{a_{i_j} \mid i_j \in I_1\}$ and $A_2 = \{a_{i_k} \mid i_k \in I_2\}$ is a solution of the equal subset sum problem of the set A.
Thus, a solution of the UCDR decision problem is also a solution of the equal subset sum problem. Conversely, we can also easily verify that a solution of the equal subset sum problem is also a solution of the UCDR decision problem. Therefore, if we can solve the decision version of UCDR then we can solve the equal subset sum problem which is NP-complete (Remark 1). Since it is easy to verify this problem is in NP, it is also NP-complete. □

1.2 Enrichment of non-LGD variants in ASD probands disrupting the selected genes and their regulatory elements.

A total of 516 ASD simplex families from SSC were recently WGS and *de novo* variants in the affected probands and unaffected sibling were predicted and validated [57]. Note that these families were selected *to be void of LGD variants based on whole-exome sequencing*. Thus, they were not part of the samples which contributed to Odin training. However, we did observe a significant number of the affected probands in comparison of unaffected siblings had non-LGD coding and non-coding *de novo* variants disrupting the coding or the regulatory elements of the genes in the inner most sphere (Supplementary Figure 1). The subset of genes in the selected 391 genes in the inner most sphere, which had a *de novo* variants disrupting their coding or regulatory elements in probands or siblings, is depicted in Supplementary Figure 2. Furthermore, we also observed the significant enrichment after removing the known SFARI high confidence and syndromic autism genes from the set of 391 genes considered.



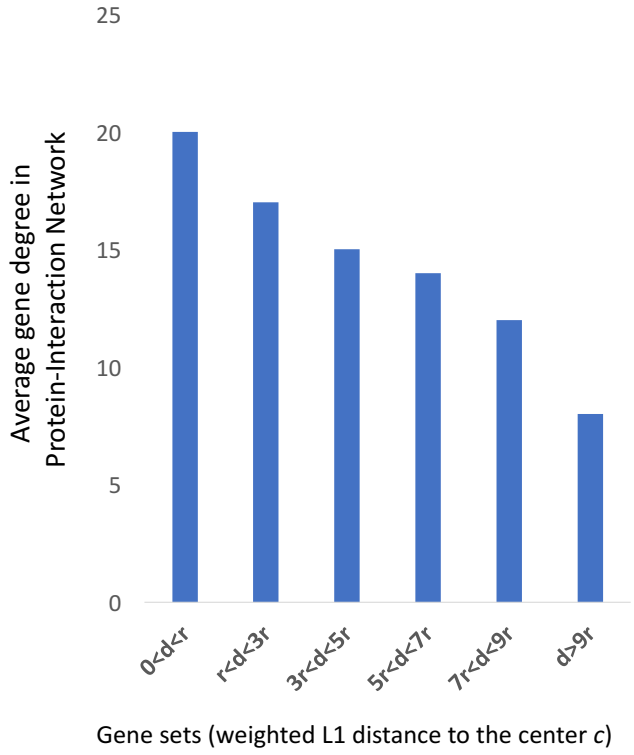
Supplementary Figure 1: Enrichment of *de novo* non-LGD variants in WGS samples disrupting coding and regulatory regions of genes in inner most sphere (total 391) in affected probands versus unaffected siblings.



Supplementary Figure 2: The non-LGD disruptive variants disrupting coding and regulatory regions of the genes in inner most sphere in ASD probands and siblings.

1.3 Protein interaction enrichment

We investigated the changes in genes degree in protein-interaction networks based on their weighted ℓ_1 distance to the center found using Odin. There is an interesting correlation between distance calculated by Odin for each gene and the average degree of that genes in protein-interaction networks (Supplementary Figure 3).



Supplementary Figure 3: The average degree of genes is higher for set of genes which are closer to the center. The center and the weighted ℓ_1 distance is learned by Odin.

1.4 Experiments details and commands

In the union of the ASD/ID datasets considered in this study (Table 1) there are a total of 684 affected ASD/ID cases/probands with LGD variants and 245 control and unaffected siblings with LGD variants. We compared the results of Odin against k-NN, SVM, and Glmnet (Lasso and Elastic-net) for predicting of ASD/ID with low false positive rate ($< 1\%$). We used a leave-one-out approach to compare these methods. We used the scores/confidence/probability outputted by each method for each prediction to control for the number of unaffected samples predicted by mistake as case (denoted as false-positive rate). The exact commands used for each program is as follows:

SVM experiments The command for training and testing used in for SVM is based on libSVM version 3.21 implementation [54]. Using the full dataset we first found the optimal parameters for "gamma" and "cost" and were set to 0.25 and 0.03125 respectively for the libSVM classifier. Then, for the LOO experiment we used the following commands in training dataset: `svm-train -b 1 -w0 5 -w1 1 -c 0.03125 -g 0.25 training-data` and in the case of test data we use the following command: `svm-predict -b 1 testing-data training-data.model output`.

Lasso and Elasticnet (Glmnet) experiments The commands used for Glmnet (lasso and Elastic-net) [55]. In training dataset we use the following command: `fit=glmnet(training-data.features, training-data.class, alpha=a` (we ran with parameters $a \in \{0, 0.25, 0.5, 0.75, 1\}$, and in the case of test data we used `predict(fit, testing-data, s=0.042645)` (the value s was calculate as *lambda.min* as instructed in https://web.stanford.edu/ hastie/glmnet/glmnet_alpha.html).

K-NN experiments We implemented the k-NN classier and tested and reported the results for k ranging from 1 to 20.

Random forest experiments We used the package randomForest in R with the following command for training

```
rf <- randomForest (training-data.features, training-data.class, ntree = 1000)
```

and the following command for testing

```
pred <- predict(rf, testing-data, type="vote")
```